REPORT RESUMES

ED 018 162 48 FL 000 663

AN EXPERIMENTAL STUDY OF THE RELATIVE EFFECTIVENESS OF FOUR SYSTEMS OF LANGUAGE LABORATORY EQUIPMENT IN TEACHING FRENCH PRONUNCIATION.

BY- YOUNG, CLARENCE W. CHOQUETTE, CHARLES A.
COLGATE UNIV., HAMILTON, N.Y.

PUB DATE 9 APR 63
CONTRACT OEC-SAE-8791
EDRS PRICE MF-$0.50 HC-$4.64 114P.

DESCRIPTORS- *AUDIO PASSIVE LABORATORIES, *AUDIO ACTIVE LABORATORIES, *FRENCH, *AUDIO ACTIVE COMPARE LABORATORIES, *PRONUNCIATION INSTRUCTION, TAPE RECORDERS, COMPARATIVE ANALYSIS, EQUIPMENT UTILIZATION, ENUNCIATION IMPROVEMENT, EQUIPMENT EVALUATION, LANGUAGE LABORATORIES, LANGUAGE LABORATORY EQUIPMENT, LANGUAGE LABORATORY USE, SECOND LANGUAGE LEARNING, AUDIO EQUIPMENT, LISTENING SKILLS, ANALYSIS OF VARIANCE, SPEECH SKILLS, COLGATE UNIVERSITY, HAMILTON, NEW YORK,

A SERIES OF SEVEN EXPERIMENTS TESTED THE RELATIVE EFFECTIVENESS OF USING FOUR TYPES OF LANGUAGE LABORATORY EQUIPMENT FEATURING INACTIVATED OR ACTIVATED FEEDBACK (IF OR AF) OR LONG OR SHORT DELAY PLAYBACK (LD OR SD) IN LEARNING TO PRONOUNCE FRENCH. AFTER PRELIMINARY EXPERIMENTATION, THREE REPLICATION EXPERIMENTS WERE CONDUCTED WITH PAID JUNIOR HIGH, SENIOR HIGH, AND COLLEGE PARTICIPANTS WHO HAD NEVER STUDIED FRENCH. PROCEDURES CHARACTERISTIC OF EACH EXPERIMENT WERE 6-DAY TRAINING SESSIONS CONSISTING OF CLASSROOM INSTRUCTION FOLLOWED BY LABORATORY PRACTICE WITH A SPECIFIC TREATMENT CONDITION AND PRE- AND POST-TESTING OF DAILY AND CUMULATIVE PRONUNCIATION MASTERY. A COMPARISON OF GROUP ACHIEVEMENTS BASED ON 24 ANALYSES OF THE EXPERIMENTS, PHONEMIC AND OVERALL PRONUNCIATION VARIABLES, AND PRE-, POST-, AND FINAL TRAINED AND UNTRAINED TESTS REVEALED THAT, DESPITE THE RELATIVELY CONSISTENT HIGH AND LOW AVERAGES FOR THE AF AND SD EQUIPMENT, NONE OF THE FOUR TREATMENTS PROVED TO BE SIGNIFICANTLY SUPERIOR TO THE OTHERS. HOWEVER, THE ANALYSES DID INDICATE FUTURE RESEARCH NEEDS, SIGNIFICANT DEFICIENCIES IN THE COLLEGE GROUP PERFORMANCE, THE USEFULNESS OF PRE-CLASSROOM INSTRUCTION, AND THE INEFFECTIVENESS OF EQUIPMENT WITH MINOR DEFICIENCIES IN SOUND QUALITY. (AB)

NDEA VI

A

ERIC

Lang

No. 8 & No.

PA48

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED018162

AN EXPERIMENTAL STUDY OF THE RELATIVE EFFECTIVENESS OF FOUR SYSTEMS OF LANGUAGE LABORATORY EQUIPMENT IN TEACHING FRENCH PRONUNCIATION

by

Clarence W. Young and Charles A. Choquette

Colgate University
April 9, 1963

FL 000 663

[redacted]

The research reported herein was performed pursuant to a contract with the U. S. Office of Education, Department of Health, Education and Welfare, National Defense Education Act, Title VI, Contract Number SAE-8791.

FOREWORD

The authors gratefully acknowledge the assistance afforded them by the following consultants and advisers.

In the fields of linguistics and language teaching: Simon Belasco, Marie-Jose Berlin, Gabriel Cordova, James Ferrigno, Alfred S. Hayes, James E. Iannuci, Ivo Malan, Stanley Sapon, and Elwyn Sterling.

In the fields of statistics and experimental design: Jack Finger, Donald L. Meyer, George E. Schlesser, and Harry Snyder.

They are particularly grateful for the willing cooperation of the high schools of Hamilton, Madison, and Morrisville, New York and of the following officers of those schools who gave generously of their time in the selection and scheduling of subjects for the experiments: Andrew L. Lane, Rodney Pierce, Conrad Ruppert, Eugene Smith, and Neil O. Wooley.

They wish also to express thanks to the many individuals who worked conscientiously and efficiently in carrying out the innumerable details of experimental procedure and statistical analysis as well as to their subjects whose friendly cooperation contributed so much to the success of the experiment.

Finally, they wish to compliment Virginia B. Young for her fine work in the arduous task of typing this report.

TABLE OF CONTENTS

Page

TEXT

I. ANALYSIS OF THE PROBLEM ---------------------------- 1

II. SEQUENCE OF EXPERIMENTS ---------------------------- 9

III. EQUIPMENT 11

IV. SUBJECTS -- 17

V. MATERIALS AND PROCEDURES -------------------------- 19
  Development and Selection of Testing and Training Procedures ---------------------------- 19
  Classroom Training ---------------------------- 24
  Schedule for Daily Training Sessions ----------- 27
  Testing and Training Materials ---------------- 28
  Scoring -------------------------------------- 30
  The Questionnaire ---------------------------- 32

VI. RESULTS AND DISCUSSIONS ------------------------ 33
  The Questionnaire ---------------------------- 33
    Results ---------------------------------- 33
    Discussion ------------------------------- 36
  The Replication Experiments(Experiments 4,5 and 7 38
    Results ---------------------------------- 38
    Discussion ------------------------------- 50
  The Continuation Experiment (Experiment 6) ----- 58
    Results ---------------------------------- 58
    Discussion ------------------------------- 59
  The Trial Experiment -------------------------- 65
    Results ---------------------------------- 65
    Discussion ------------------------------- 65
  General Discussion and Suggestions for Further Research ---------------------------- 67
    Activated vs. Inactivated Feedback ---------- 67
    Delayed Playback ---------------------------- 68
    Standardizing the Testing of Pronunciation -- 71
    Age Level and Pronunciation Aptitude -------- 72
    Types of Research on Pronunciation ---------- 74

VII. SUMMARY -- 76

APPENDIXES

I. TESTS AND TRAINING PROGRAMS ---------------------- 80
II. QUESTIONNAIRE ---------------------------------- 92
III. ANALYSIS OF OPEN-ENDED ITEMS ON QUESTIONNAIRE ----- 93
IV. SUPPLEMENTARY TABLES ---------------------------- 100

AN EXPERIMENTAL STUDY OF THE RELATIVE EFFECTIVENESS OF FOUR SYSTEMS OF LANGUAGE LABORATORY EQUIPMENT IN TEACHING FRENCH PRONUNCIATION

by

Clarence W. Young and Charles A. Choquette

I. ANALYSIS OF THE PROBLEM

The past fifteen years have seen a remarkable growth of language laboratories for the teaching and learning of a second language. The essential teaching device of the language laboratory is to permit students to hear and often to mimic or reply to recorded utterances in the second language. One of the major advantages of this teaching method is thought to be the better learning of pronunciation, since recordings can be made by native speakers of the second language, and the student may listen to and attempt to imitate their pronunciation in continuous practice sessions. Another assumed advantage is that, by presenting the recordings through headphones, the student may be expected to hear the sounds more clearly than in the average classroom.

Three systems of record playing equipment are in general use: (1) the so-called audio-passive system, in which the student imitates recorded material while listening through headphones alone; (2) the audio-active system, in which the student hears his own voice electrically amplified as he speaks; (3) equipment which enables the student to record his imitation of a master sample, then play back both master sample and his imitation thereof for self-evaluation purposes.

In commercially available versions of the third system, there is frequently a relatively long delay between actual recording and subsequent playback. Some teachers have felt that a drastic reduction in this delay, permitting the student to evaluate his response almost immediately, would be still more effective. Such a short delay playback arrangement may be designated as a fourth possible system.

Audio-active systems are somewhat more expensive than audio-passive systems, and playback systems are considerably more expensive, since they involve the purchase of recording equipment for each student. Such recording equipment may also be used to permit each student to select his own practice program; but where, as is usually the case, a single program is presented to an entire class, the only possible advantage of the more expensive systems over the simple audio-passive system is better self-monitoring with regard to pronunciation. Presumably the playback systems afford the student a better opportunity to hear and correct his errors in pronunciation. There are obviously important economic reasons for determining whether or not the more expensive systems are actually more effective. But, of course, there are equally obvious pedagogical reasons for determining which of the four possible systems is the more effective teaching device.

The aim of the present study has been to determine which, if any, of the above four systems is most effective in teaching French pronunciation to beginning students during the first stages of learning.

The terminology ordinarily employed to distinguish between these systems refers to differences in the mechanical arrangements. The nature of the problem will be clarified if the terminology is more closely related to the effect on the student. Thus viewed, the problem is one of discovering the optimum system for correcting wrong pronunciation and for being reinforced for approximately right pronunciation.

Both the audio-passive and audio-active arrangements provide an _immediate feedback system_, which tells the student, while he is speaking, what auditory pattern has resulted from the action of his speech muscles. We may distinguish between these two systems by calling one the _inactivated feedback system_, or _IF system_ and the other the _activated feedback system_, or _AF system_.

In both the IF and the AF systems, the acoustic input to the cochlea is conducted both through the bones of the skull and through the system for receiving external sounds; that is, the meatus, tympanic membranes, ossicles, and oval window. The first type of input may be termed internal, the second external. The chief difference between inactivated feedback and activated feedback is one of the ratio between internal and external input. With IF, the relative amount of external input is reduced, although it is never eliminated, since it is always possible to hear external sounds while wearing the headphones. The amount of reduction depends on the type of headphone used. Heavily padded headphones will effect a greater reduction than unpadded phones. Hence, hearing one's voice in the IF condition is more like ordinary hearing if headphones are not padded. With the AF condition, the ratio depends on the amount of gain applied to the headphones. As the gain is increased, the proportion of external input is increased. Relative to the intensity of vocalization, the total intensity of input is also increased.

With respect to their effects on the student, therefore, the IF condition is best described as "reduced external feedback" and the AF condition as "variable external feedback."

Subjectively -- at least according to our introspections -- the localization of the sound varies with the ratio of external to internal input. In ordinary speech -- unless there is a strong echo in the room -- the voice appears to be vaguely localized immediately in front of the face. Upon covering the ears with the hands, the localization moves inward; that is, the sound seems to be localized within the head. In the IF condition with unpadded headphones, the sound is only slightly inward. Actually, the sound is more like the voice in ordinary speech than the voice with ears covered by hands. With the AF condition at low gain, the localization moves back to the position characteristic of ordinary speech, but as the gain increases, the sound is heard more and more in the headphones themselves. As the intensity reaches the level of unpleasantness, the sound seems to penetrate the ears and enter the head. As the sound "moves" from its position in the IF condition to the ordinary position, then to the headphones, and then into the ears and head, its ability to compete with other sounds for attention seems to increase. This may be a function of intensity, although low speech sounds may be heard in the headphones with what seems to us to be an increased salience. Similarly, as the sound moves from the glottal region to the forehead and to the headphones, it seems to grow clearer, and sound differences, or variations in intonation, seem to be more readily discriminated. As the gain is further increased and intensity reaches higher levels, however, clarity decreases rapidly. As might be expected, when the sound is localized in the region of the face, it sounds more like one's own voice than in any other position.

With relatively low gain, the activated feedback condition produces an impression similar to ordinary speech conditions. When, with higher gain, the voice is heard in the headphones, it becomes more like the model that the student is imitating, and it may actually provide conditions for discrimination superior to those in ordinary speech. Actually, nothing is really known about this, and we do not know what level of gain, if any, is optimal for activated headphones. At present, the attempt is made to set the gain so as to produce maximum comfort for the student. We know little of just how different the student's experience usually is under the IF and AF conditions. Indeed, it is not beyond the range of possibility that, under ordinary laboratory conditions, the distraction and masking provided by the sound of neighboring voices may eliminate whatever differences might otherwise exist between the effectiveness of the activated and inactivated conditions. Because the microphones pick up other sounds than those of the speaker's voice, this distraction tends to be greater for the activated condition unless unidirectional, close-speaking microphones with heavily padded headphones are used.

Obviously, there are many specific ways in which both the inactivated and activated headphones may be used, and the differences between them exist on a quantitative continuum. Psychologically, as the student experiences them, inactivated feedback without padded headphones may be more like activated feedback with unpadded headphones and low gain than the latter is like activated feedback with heavily padded headphones and the gain turned to the point where the locus of sound is entirely in the headphones and ears. And the latter condition may be less like the condition of ordinary speaking than either of the two former conditions.

There are many possible arrangements for both IF and AF, not all of which have been analysed above. It might well be that the most effective inactivated system would be superior to certain less effective activated systems, and vice versa. Or it might be that any reasonably good immediate feedback system is as effective as any other. There seems to be a tendency to assume that the AF condition must necessarily be superior to the IF condition. But the following factors may be suggested which might make for superiority in an IF system:

(1) It might be that lack of self-consciousness about pronunciation is a positive factor in learning to pronounce through mimicry. The IF system, since it does not require speaking into a microphone and does not emphasize the sound of the speaker's voice, might tend to reduce self-consciousness.

(2) The need to speak into a microphone complicates the situation. This makes for some distraction. There is also the need for adjusting gain to produce an optimal comfort. This can create problems and distractions.

(3) The possibility cannot be excluded that bone-conduction hearing gives a better cue to the accuracy of pronunciation -- for some phonemes at least -- than internal plus external input. The psychologist author of this report, for example, believes that he can perceive diphthongization of vowels more readily with ears covered than without.

The possible advantages of AF are, of course, more obvious. One clear advantage is that, with an activated hookup, the gain can be adjusted so as to produce many degrees of external input, from very low to the highest tolerable. In principle, it would be possible to find the optimal amount of external input and use that. Since an AF arrangement can reproduce the conditions of IF, the latter can be superior to the former only because of its greater simplicity and lower expense.

The above analysis is not intended to settle our problem in advance, but rather to demonstrate the impossibility of making a priori judgments by pointing to a few of the possible factors that might make for differences or for absence of differences in effectiveness between the IF and AF conditions.

Correlative with our designation of the two immediate feedback conditions as "inactivated feedback," or IF, and "activated feedback," or AF, we shall designate the two playback conditions as "long delay," or LD, and "short delay," or SD. Each of these conditions involves recording the student's voice as he imitates the model. Hence some of the student's learning is achieved in the condition of immediate feedback, whether IF or AF. The effectiveness of any particular course of training with playback must, therefore, depend in part upon the particular system of immediate feedback with which it is combined. There are not only many possible immediate feedback arrangements, but there are many ways in which playback may be combined with immediate feedback. Obviously, a whole series of delay periods, from a fraction of a second up to several days can be introduced. It is possible, for example, that it would be more effective to delay playback until just before a practice session to give the most useful information on the aspects of the student's pronunciation that need to be improved. A fairly obvious weakness of any system of training using playback is the fact that time spent listening to playback is subtracted from time that might be spent in active practice. However, the proportion of time spent in practice to that spent in listening may be widely varied. For example, if forty minutes were to be spent in practice three times a week, it would be possible to record only the last four minutes and then to play it back at the beginning of the next practice session. This might enable the student to estimate his most characteristic successes and failures and provide him with more purposeful goals for the ensuing practice session. Or short delay practice might be made to occupy the first few minutes of each session. Countless other temporal arrangements are obviously possible.

The foregoing discussion should make it clear that no experiment could possibly be set up which could be certain of providing a definitive test of the relative effectiveness of the four equipment systems unless it could be known in advance which particular arrangement is optimal for each system. We do not have this knowledge or anything approximating it. Furthermore, our investigation has been limited to the first few hours of training. A training system might be shown to be highly effective for the first approach to language training, yet in the long run it might be of little value. On the other hand, a system might show little advantage at first, but on the basis of a slight superiority in some respect, its cumulative advantage might be great.

It follows that the mere finding of statistically significant differences between the treatments would not constitute a definitive determination of relative superiority, since changes in procedure or specific kinds of equipment might reverse the findings. If we let A, B, C, and D stand for the particular form of IF, AF, LD, and SD we might actually use, and Ax, Bx, Cx, and Dx stand for any other possible forms of each condition, the finding that A is significantly superior to B, C, and D would not prove it superior to Bx, Cx, and Dx, nor would it bring proof of the superiority of Ax to B, C, and D or to Bx, Cx, and Dx.

The same restrictions on generalization apply with respect to the group of subjects selected for a given experiment. A method that might be superior for one set of subjects might not be superior for all other sets. Furthermore, different kinds of program, different schedules of study, as well as different amounts of total time spent might result in differences in the effectiveness of the four conditions.

This kind of difficulty is faced by most experiments in the behavioral sciences, and no responsible behavioral scientist is likely to come to a final generalized conclusion on the basis of a single experiment. An experiment gets us better acquainted with a certain area of phenomena, narrows the range of uncertainty, and suggests strategic approaches to further experiments which will further narrow the range of uncertainty. At any stage in the process of narrowing this range, practical decisions must be made on the basis of best estimates of the true relationships among variables.

For example, prior to our experiment, there could be no basis for rejecting the hypothesis that the use of a short delay system is by far the best method of initiating the teaching of pronunciation. If the SD treatment had turned out to be markedly more effective than the others, this hypothesis would have received strong confirmation, and the range of uncertainty would have then been narrowed. Practically, such a finding would lead to decisions to produce short delay systems commercially and to test their usefulness more widely. There would still remain the following questions:

(1) Does the short delay system maintain its superiority over a period of time?

(2) Are there methods of employing the other equipment systems in a different way than they were employed in our experiments which make them as effective as or more effective than the SD system?

(3) What temporal combination of short delay with immediate feedback systems is most effective?

(4) What particular short delay arrangements are most effective?

As a matter of fact, our experiment has not provided the finding hypothesized above. Hence, it has narrowed the range of uncertainty in a different way and raised a different set of questions.

To summarize: The general problem approached in the present experiment contains too many variables to be tested in a single study, and the experiment cannot be expected to provide a final and certain answer to the practical problem of the best kinds of equipment systems and the best manner in which to employ them. The aim of the experiment is exploratory: To get some measure of the relative effectiveness of the systems under circumstances designed to give each one an opportunity to display its merits.

Various considerations lead to contrary theoretical assumptions concerning the probable effectiveness of the four systems. It might be assumed that the immediate feedback conditions should be favored because immediate reinforcement or knowledge of results is typically found to be superior to delayed reinforcement or knowledge of results in learning situations. The playback conditions, on the other hand, might be favored by the greater clarity or certainty of the knowledge of results. Under immediate feedback, knowledge of results is received under the strain of actually speaking the utterance, and the attention may thus be distracted from the task of discriminating success and failure. The short delay system used in our study played back the student's voice within a second and a half after he began his utterance. Our hypothesis that this might be especially effective was as follows: As the student speaks the utterance, he receives immediate knowledge of results. With less than a second's delay, he receives a presumably clearer knowledge of results. This, we hypothesized, might provide him with a doubly strengthened basis for eliminating errors and establishing correct habits. He might immediately vary his mode of pronunciation and discover exactly what motor patterns produced the best results, and his judgment of these results might be based on better listening conditions than are provided by immediate feedback alone.

Under the long delay condition, there is no opportunity to judge the success of a particular motor pattern of vocalization and correct it, since when the student hears his utterance, he has no way of remembering how he made it. Long delay, however, affords a particularly good opportunity for the student to observe the difference between his pronunciation and that of the model. Short delay may be described as an attempt to secure the advantages of both immediate feedback and playback: relatively clear and undistracted information combined with a relatively short interval between the response and the knowledge of results. A possible handicap for short delay is the fact that the student receives both immediate feedback information and playback information in short succession relative to a single response. It might be that making judgments based on two types of information could actually lead to some confusion.

In spite of the possibility of this handicap we were inclined, prior to running our experiment, to expect great things of the SD condition, since it appeared to be favored both by short delay between response and knowledge of results and by greater clarity in knowledge of results. A major consideration in selecting the type of experiment we chose was to test the possibilities of this new arrangement for playback. Otherwise a comparison of the three other treatments under the condition of course teaching might have seemed preferable. In the absence of already established programs for short delay playback, however, it was necessary to make a test involving a relatively short training period in a special experimental situation.

As a by-product of our study, we sought to compare the rates of learning of junior high school students (seventh and eighth grades), senior high school students (ninth, tenth, and eleventh grades) and college men. This was to test the common belief of language teachers that younger students are more apt in the learning of pronunciation. We also studied the problem of reliability and validity in the measurement of pronunciation.

II. SEQUENCE OF EXPERIMENTS

The experiments were carried out with subjects who had no training in or experience with the French language and who were paid for their service. Ordinary laboratory teaching conditions were duplicated to the extent that the training occurred in the language laboratory with groups of subjects. To restrict the experiments to the skill of pronunciation alone, the subjects had no knowledge of the meaning of the utterances they pronounced, nor were they shown the printed words until the experiment in which they served was completed.

Seven experiments were run in the course of the study. The first two of these were designed simply to develop testing and scoring procedures. The third was a trial experiment testing the effects of the IF, AF, and LD treatments only, since the equipment for the SD treatment had not yet been constructed. Thirty-four junior and senior high school students served as subjects in this third experiment. It was followed by a series of four experiments testing the effects of all four treatments in which the testing and training material was modified as a result of the experience gained in Experiment 3.

In Experiment 4, 28 senior high school students served as subjects. In Experiment 5 there were 28 college students, but one of them, in the SD group, dropped out after the third day. Twenty-eight junior high school subjects served in Experiment 7. Experiment 6 was a continuation of Experiment 5 with the same group of college subjects. It was designed to test the effects of a longer period of training.

Experiments 3 through 7 constituted the entire series of Training Experiments. They are summarized in Table I. Experiment 3 will be called the Trial Experiment and the remaining four, the Main Experiments. Since Experiment 6 was a continuation of training with the group of college subjects in Experiment 5, it will be termed the Continuation Experiment. Since Experiments 4, 5, and 7 were exact replications of one another except for the subjects, they will be called the Replication Experiments. Essentially the same schedule was followed in all five Training Experiments. Except for Experiment 3, an Aptitude Test was given the first day. In Experiment 3, the subjects were practiced in handling the machines and mimicking French utterances for three days and then given an Aptitude Test the fourth day. Following the Aptitude test in all five experiments, the subjects were divided into treatment groups and given six Training Sessions on each of six days. On the Day following the six training days a Final Test, identical in content with the Aptitude Test was administered.

TABLE I

SUMMARY OF TRAINING EXPERIMENTS, SHOWING TYPE OF EXPERIMENT, NUMBER OF EXPERIMENT, NUMBER OF INTRODUCTORY, TESTING, AND TRAINING DAYS, AND TYPES OF SUBJECTS.

Key for "Number of Days" columns:
IAT: Introduction and Aptitude Test
TS: Training Sessions
FT: Final Test

		No.Days			
Type of Experiment	Exp.No	IAT	TS	FT	Type Subject
Trial Experiment	3	4	6	1	Jr.-Sr. H.S.
Main Experiments:					
Replication Experiment	4	1	6	1	Senior H.S.
" "	5	1	6	1	College
" "	7	1	6	1	Junior H.S.
Continuation "	6	1	6	1	College

(The experiments are numbered in the order in which they were carried out.)

III. EQUIPMENT

Space Arrangements in Laboratory

The experiment was performed in a laboratory containing 34 semi-isolated booths in five ranks, seven booths in each rank except the first which contained six. The control center was located across the front of the laboratory, separated from it by a partition composed of cinder blocks to the height of four feet with plexiglas continuing sixteen inches above.

The laboratory outside the control center was approximately 24 feet long, 19 feet wide, and 7 feet 6 inches high. The ceiling was covered with sound-absorptive 3/4 inch acoustical tile. In each booth, the microphone and recorder stood on a formica-topped table 24 inches deep. The booths were 27 inches wide, separated from one another by half-inch plywood partitions. The partitions were 23 inches high and 29 inches long, so that they projected 5 inches beyond the edge of the table where the subject sat. The backs of the booths were of plexiglas. Extending from the back, an 18 inch wide strip of plexiglas covered the top of each booth.

This arrangement provided a considerable degree of isolation, and at the same time allowed the experimenter at the control center to watch and communicate with the subjects. Communication from the experimenter was secured with a microphone fed into the machine that played the program tape, and communication from subjects was achieved through hand signals.

List of Equipment and Specifications

Wollensak T-1500 tape recorders (used for IF, AF, and LD)
Wollensak T-1515-4 tape recorders (used for SD)
Wollensak tape recorder microphones (subjects' microphone)
Revere T-202 tape recorders (used for program tapes and producing scorers' tapes)
Ampex 351 tape recorders (used in recording program tapes and scorers' tapes)
Electro-Voice 664 cardioid microphone (used in recording program tapes)
Heath EA-1 audio amplifiers (used for mixing inputs and driving earphones in short delay mechanism)
Shure TR5B-J magnetic recording head (used for delayed playback pickup in short delay mechanisms)
General Electric UPX-003B pre-amplifier (used for pre-amplification of Shure recording head output)
Viking AS-75 amplifiers (used to provide "activation" of subjects' earphones)
Military HS-33 600-ohm magnetic earphones (used as subjects'

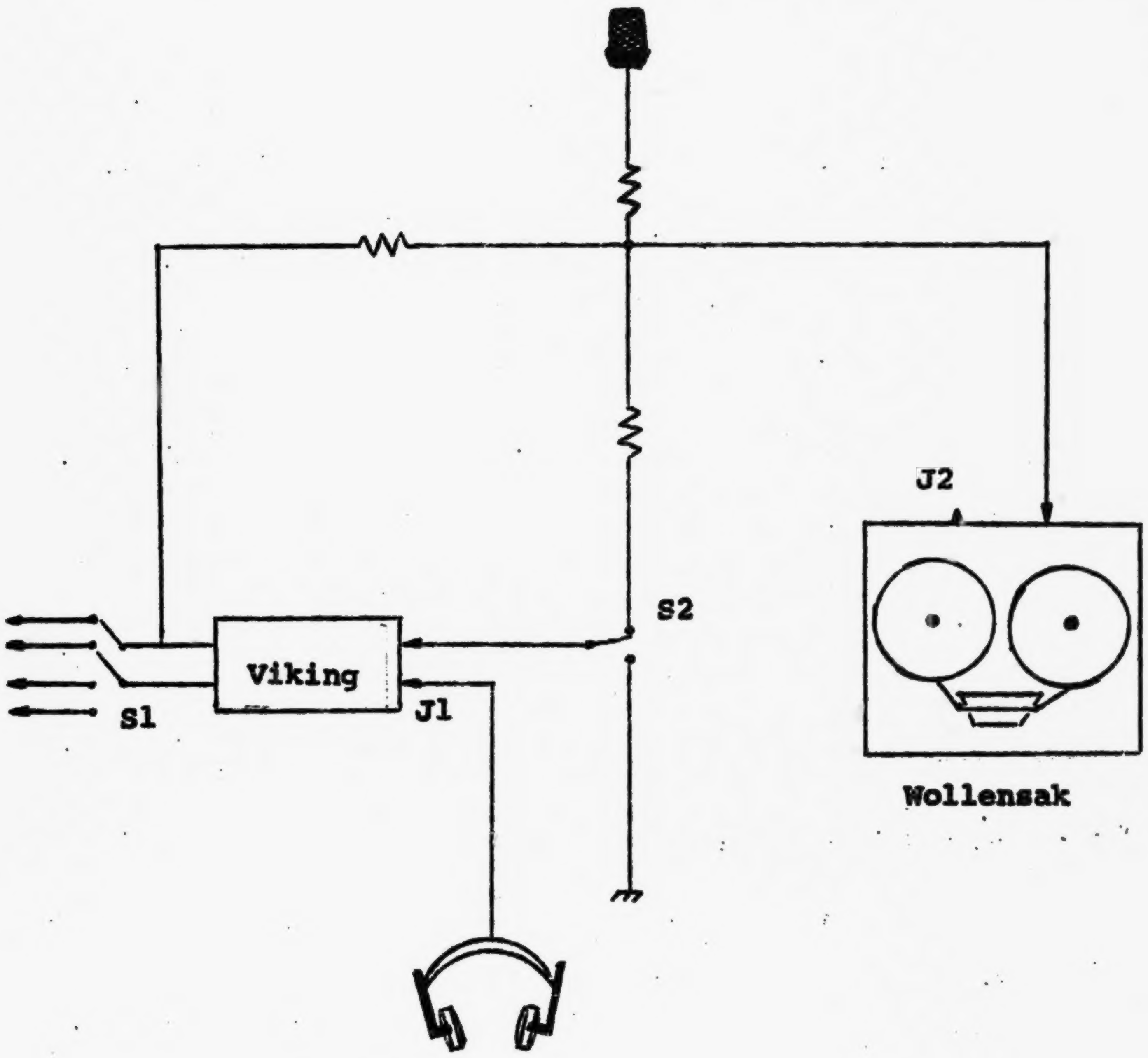

Fig. 1. Diagram of individual booth wiring. (See text for description.)

earphones)
Scotch 311 Tenzar recording tape (used as program tapes and for SD recording)
Scotch 190 Acetate recording tape (used for subject recordings for IF, AF, and LD)
Scotch 111 Acetate recording tape (used for scorers' tapes)

Frequency response:	listening--60-11,300 cps within 2 db. recording--60- 7,500 cps within 2 db. (high frequency response limited by microphone)
Noise and hum:	minimum of 50 db. below saturation recording level and at least 55 db. below normal playback level for IF, AF and LD. 22 db. below normal playback level for SD.

Distortion: combined harmonic and IM below 2.5% of total sound energy within specified frequency range

Wow and flutter: 0.23% (deviation less than +.03% for any machine)

Tape speed: standardized at 7½ ips

Wiring System to Booths

The audio transmission system used 600-ohm balanced lines throughout. Greater than 55 db. isolation between any two pairs was provided through grounded shielding of floating balanced pairs. Earth ground was established at one point only and all shield drain wires and jacks were isolated. Level on the system was not in excess of +4 VU where zero VU is equivalent to one milliwatt across a 600-ohm resistive load. Major wiring was provided by Belden 8766 cable.

Booth Wiring

The following description applies to the wiring for the Main Experiments. In the Trial Experiment, the signal from the student's microphone was fed through the Wollensak amplifier for purposes of activation, with a resulting mismatch between the microphone and tape recorder input. This was judged to produce a sound that was inferior to both the sound from the program tapes and from the LD playback. To equalize conditions for activation, Viking AS-75 amplifiers were introduced.

Individual booth wiring is represented in Fig. 1. Switch S1 provided two alternative input sources both of which terminated at the control center. For AF, the chosen input fed simultaneously the Viking amplifier and the tape recorder input. The Wollensak microphone also simultaneously fed both inputs through a resistive balancing network. In this situation, the student heard both the program and his own recording voice. For LD and for tests the Wollensak recorded both the program source and the subjects'response. By shunting the Viking microphone input with switch S2, activation was eliminated. For LD playback, the headphones were removed from the Viking (J1) and inserted in the Wollensak external speaker jack (J2).

Thus, through combinations of the above conditions, the IF, AF, and LD situations were all made available. The booth wiring system was the same in the SD booths,and the SD playback could be switched in or out.

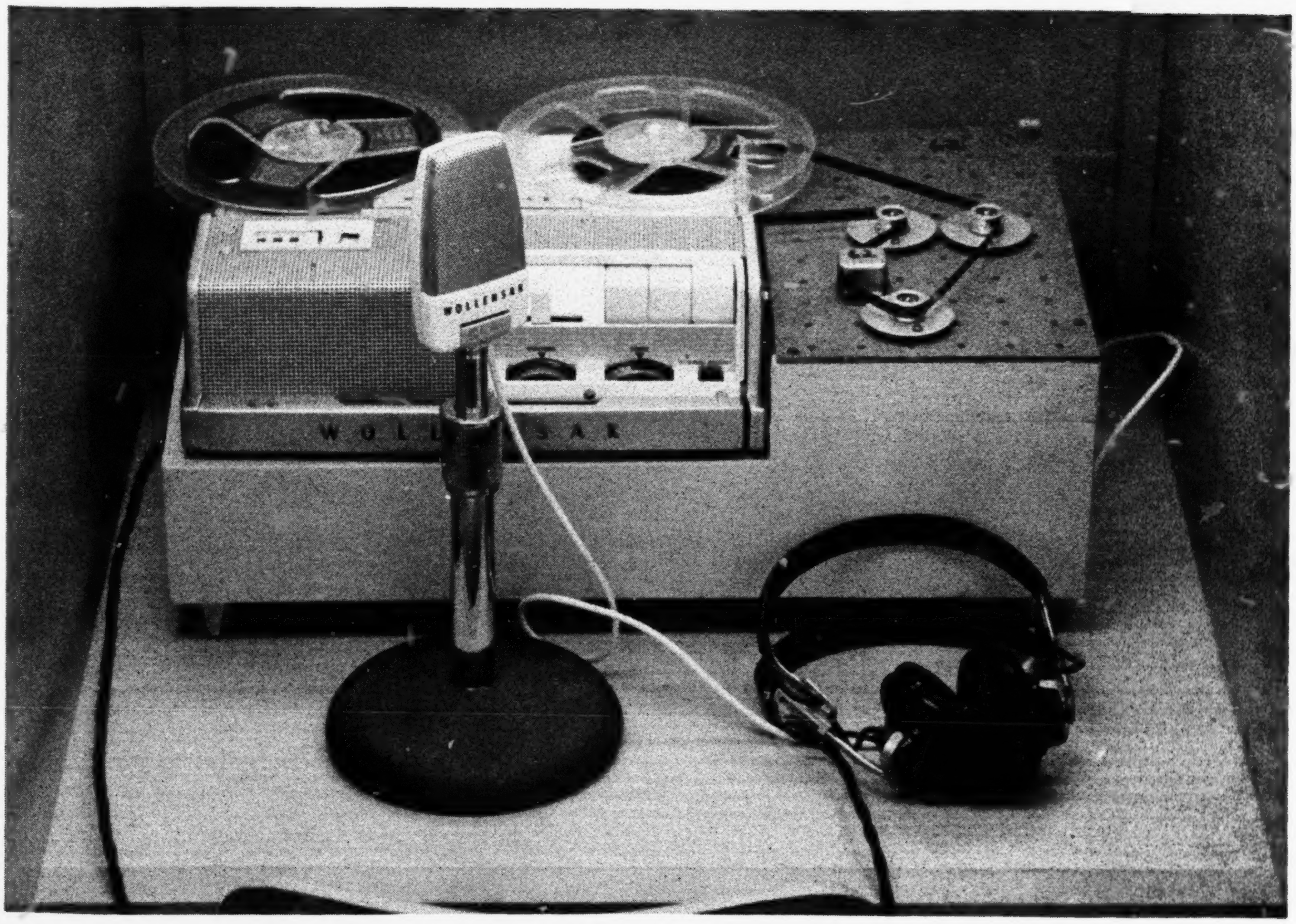

Fig. 2. Short delay playback system.

The Short Delay Playback Unit

The short delay playback system is shown in Fig. 2 and diagrammed in Fig. 3. The basic recorder was the Wollensak T-1515-4. Attached to its right side was a perforated masonite plate on which was fixed a series of tape-directing rollers to carry the tape past the playback recording head. A Shure TR5B-J recording head (RH) for playback pickup was mounted along the tape route and provided the delayed playback pickup. The playback head output was fed through a General Electric UPX-003B pre-amplifier (GE) (modified for standard NAB tape head equalization) to a junction containing the two other sources (program from control center and student's voice from Wollensak output). This signal combination was mixed, equalized and then fed to the crystal input of the Heath EA-1 audio amplifier (mounted directly beneath the perforated masonite plate). The EA-1 output, terminated with an 8-ohm resistive load, then finally drove the subjects' HS-33 earphones at an adjustable gain. The subjects' voice was recorded through the Wollensak recording head. The tape distance between

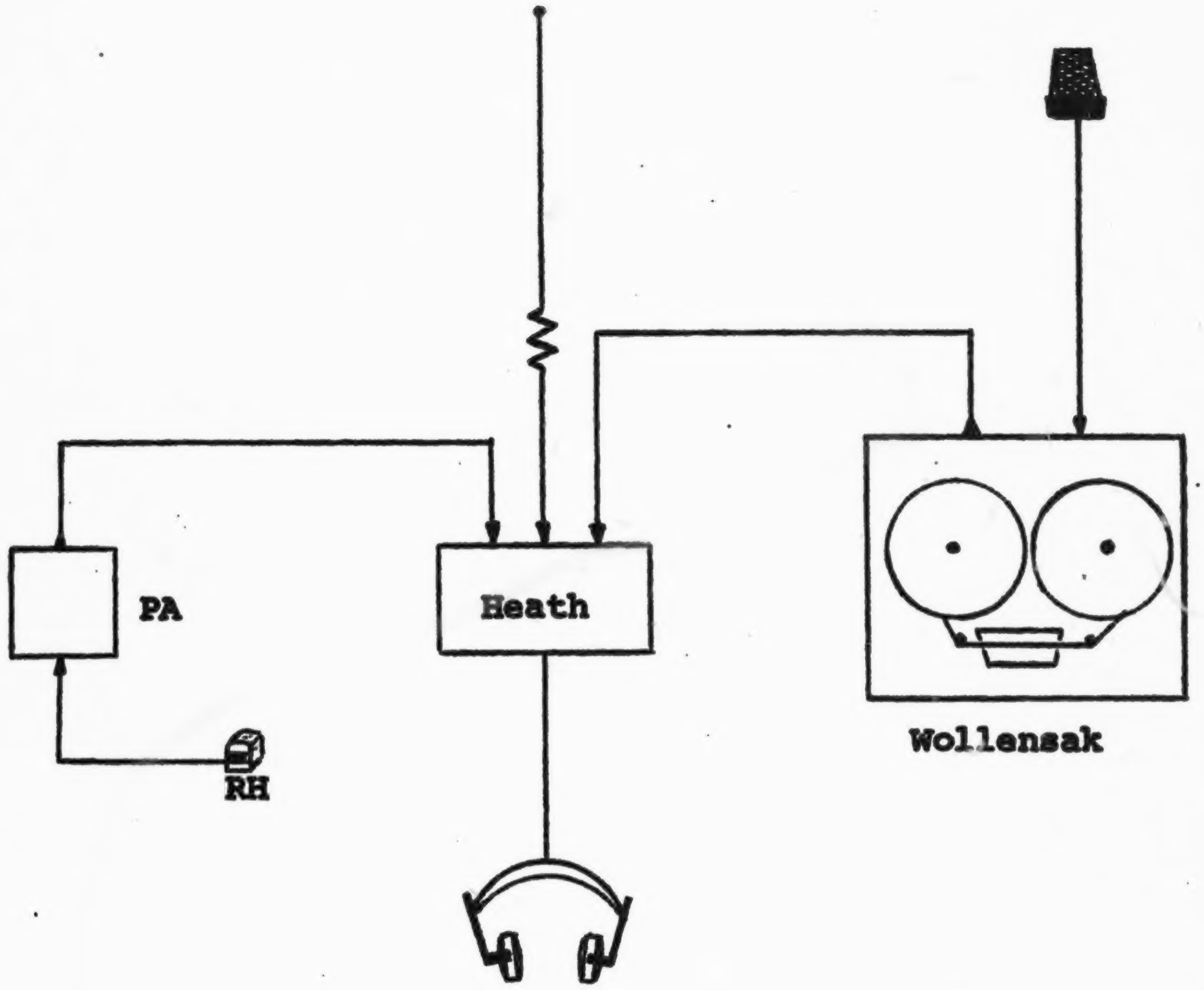

Fig. 3. Diagram of Short delay playback system. (See text for description.)

P.A: Pre-amplifier
R.H: Recording head

the recording head and the playback head was 11 inches, thus allowing a 1½ second delay between the beginning of a subject's utterance and the playback of the utterance.

The quality of the sound system for the IF, AF, and LD groups was judged to be "the best ever heard in a teaching laboratory", by a national authority on sound systems who possessed a wide range of experience. The quality of the SD system was judged by the experimenters to be not as good, partly because of the relatively lower elevation of signal over noise. This difficulty appeared even when the playback was switched out. The quality was good enough, however, to permit clear discrimination of speech sounds at all times. To accommodate to other activities of the subjects, it was necessary to begin Experiment 4 before the SD equipment had been fully tested. Throughout the first four days of the experiment, breakdowns in the SD equipment occurred which made it necessary to shift subjects to a standby equipment or to make adjustments during the experiment.

In brief, the experimenters did not succeed in perfectly equating equipment conditions between the SD group and the others, and this may have produced some handicap for the SD condition.

IV. SUBJECTS

The college students in Experiments 5 and 6 were secured from the student body of Colgate University, an all men's school. The junior and senior high school students came from nearby high schools. The junior high school students for the trial experiment came from Madison High School, the senior high school students from Hamilton High School. Senior high school students from Morrisville High School served in Experiment 4 and junior high school students from the same school in Experiment 7.

It was impossible to get a representative sample of students, since it was necessary to accept almost all those who volunteered. The volunteers tended to come from the more able and serious students in all schools. The tendency to get the more conscientious students was a distinct advantage, since it was necessary to gain cooperation for a rigidly programmed and rather artificial procedure which involved a somewhat monotonous series of repetitions. Most of the senior high school students were girls because athletic activities made it impossible for the boys to participate.

All prospective subjects were given Part II of the Carrol-Sapon Modern Language Aptitude Test, which was chosen as a preliminary test of pronunciation aptitude on the basis of the statement in the Manual[1] that "it tends to correlate highly with the ability to mimic speech sounds and sound combinations in foreign languages."

In Experiment 3, subjects who scored below 17 on the MLAT tended to speak with such low voices that their utterances could not be scored, and they often failed to respond. For the Main Experiments, therefore, prospective subjects were given a voice test, and those failing to speak loudly enough as well as those scoring below 17 on the MLAT were rejected. This resulted in the rejection of five prospective subjects.

It was necessary to drop some subjects from the statistical analysis of Experiment 3 as explained in the section on Development and Selection of Testing and Training Procedures. The result was that only twenty-four of the original thirty-four subjects in Experiment 3 entered into the experimental analysis.

---

[1] Carroll, John B. and Sapon, Stanley M. Manual for Modern Language Aptitude Test. New York: The Psychological Corporation, 1959.

TABLE II

SUBJECTS IN EXPERIMENT 3 BY SCHOOL PLACEMENT, TREATMENT GROUP, SEX, AND SCORE ON THE SECOND PART OF THE MLAT.

	IF		AF		LD		Dropped*	
	Sex	MLAT	Sex	MLAT	Sex	MLAT	Sex	MLAT
Junior High School	M	17	F	18	F	22	F	19
	M	20	F	19	F	21	F	16
	F	21	M	23	M	20	F	15
	F	24	F	22	M	18	F	13
	F	21	M	18			F	23
Senior High School	F	29	F	29	F	28	F	18
	F	24	F	17	F	23	F	24
	F	17	F	20	F	18	M	21
					F	24	F	17
							F	21
MLAT Mean:		21.6		20.8		21.8		18.7
MLAT SD:		3.88		3.37		2.74		3.37

*Dropped: Not used in statistical analysis.

Table II shows the sex, MLAT scores, school placement and experimental treatment of all subjects in the Trial Experiment. Table III gives the same information for the four Main Experiments. It will be noted that, except for the college group which was all male, the treatment groups were approximately equalized with respect to the number of subjects of each sex, although only in the junior high school group of Experiment 7 was it possible to get an equal number of males and females in the entire experimental group.

TABLE III

SUBJECTS IN MAIN EXPERIMENTS, BY EXPERIMENT, TREATMENT GROUP, SEX, AND SCORE ON THE SECOND PART OF THE MLAT.

	IF		AF		LD		SD	
	Sex	MLAT	Sex	MLAT	Sex	MLAT	Sex	MLAT
	F	18	F	17	F	29	F	26
	F	25	F	25	F	25	F	25
Experiment 4:	F	26	F	22	F	26	F	22
	F	19	F	24	F	18	F	26
	F	24	F	23	F	24	F	20
	M	22	F	27	M	21	F	20
	M	27	M	25	M	18	M	19
MLAT Mean:		23.3		23.0		23.0		22.6
MLAT SD:		2.96		3.2		3.96		2.82
	M	30	M	30	M	29	M	30
	M	27	M	27	M	28	(M)*	(26)*
Experiments 5 and 6:	M	17	M	24	M	25	M	25
	M	24	M	24	M	24	M	24
	M	22	M	29	M	22	M	23
	M	21	M	21	M	22	M	21
	M	20	M	19	M	19	M	20
MLAT Mean:		23.0		24.9		24.1		23.8
MLAT SD:		4.09		3.76		3.27		3.24
	F	24	F	27	F	29	F	23
	F	23	F	22	F	25	F	23
Experiment 7:	F	19	F	19	F	19	F	21
	M	20	M	20	F	17	F	18
	M	25	M	25	M	24	M	27
	M	24	M	24	M	17	M	21
	M	17	M	17	M	19	M	19
MLAT Mean:		21.7		22.0		21.4		21.7
MLAT SD:		2.82		3.30		4.28		2.76

*This subject dropped out after the third day of Experiment 5. The Mean and SD do not include this subject.

V. MATERIALS AND PROCEDURES

Development and Selection of Testing and Training Procedures

Considerable preliminary investigation preceded the final arrangement of testing and training procedures. The aim was to discover the kind of testing and scoring methods that would provide reliable and valid measures of pronunciation, to find the most effective ways of quickly adapting naive subjects to the laboratory situation, and to determine the kind of programs that would be best adapted to the subjects' capacities for improving phonological accuracy over a short period of time.

In several respects the capacities for adjustment of junior high school students were found to be inferior to those of senior high school and college students. Tests of procedure were, therefore, confined largely to individuals in the younger group, with three to six individuals not later participating in the experiments being employed in each test. Some of the final adjustments of programs and procedures were made on the basis of experience gained in Experiment 3, the Trial Experiment.

The methods finally developed were adapted to the limitations of the least able junior high school students, and the following description of limitations applies to them, although in some cases the same limitations might also be found in the older groups.

After attempts to train both younger and older groups to adjust the gain on their machines, start them, stop them, and rewind the tapes, it was decided to confine the subject's control of the machines to the function of stopping them only. This was necessary because a complete record of each student's test recordings was required for each laboratory session, and it was essential that all students work under comparable conditions for hearing both their own voice and that of the model. A single error on the part of a student in the operation of his equipment could render his performance non-comparable with the others or destroy the data he provided for an entire session.

Each experimental group, therefore, was furnished with a proctor who was highly experienced in the handling of laboratory equipment. The proctors stood at all times at the ends of the rows. Prior to every test and practice session, material was played on the master tape so that subjects could indicate by raising their hands whether or not they were receiving properly. Whenever failures occurred, the proctors immediately corrected them or, if necessary, called on the technician, who also served as one of the proctors. During the warm-ups preceding each test, the proctors passed behind each student to make sure that the flicker light on his machine indicated the proper gain for recording.

The experience gained in the Trial Experiment was of considerable value in indicating the degree of care on the part of the proctors required to keep all equipment operating properly at all times. Occasionally a machine failed to record during a test or recorded so poorly that the material could not be scored. This, along with recording failures caused by inadequacies in the subjects made it necessary to drop several subjects from the statistical analysis.

Failure to record was due to failure of the switch to make the proper contact when the "Record" key was pressed. This difficulty was overcome by the following procedures:

1. Before the machines were started, the power was turned off at the master switch. Each proctor then passed along his row and snapped down the keys with a firm thrust. The power was then turned on, and ten seconds later the master tape was started.

2. After each test, the proctors wound the tapes back a short way and then played the last utterance or two to make certain of the recording. In the rare cases when recording failed, the student was retested. Although this undoubtedly introduced a practice error, the amount of additional practice was slight compared with that provided by the Practice Sessions and warm ups. Furthermore, the few cases which did occur were about evenly balanced for the four experimental conditions.

3. Whenever a machine failed to record, it was removed for servicing and another one put in its place.

In Experiment 3 some unscorable recordings of utterances were produced because a subject spoke in a weak voice or moved his mouth too close to or too far from the microphone, or because the gain setting was poor for a particular test. The method of overcoming the first difficulty through careful selection of subjects has already been described. The second was overcome by requiring all subjects to place the microphone at right angles to the mouth and speak into it as it barely touched the cheek. Constant vigilance on the part of the proctors maintained adequate gain settings throughout the Main Experiments. As a result of the above precautions, the number of unscorable utterances for the Main Experiments was reduced to a fraction of one percent.

The preliminary experiments appeared to give rather clearcut indications that the optimum number of repetitions for a single utterance was two. Generally the subjects improved their mimicry on the second presentation of an utterance, but frequently the third repetition was not as good as the second. Furthermore, the subjects themselves said they wanted a second chance at an utterance, but did not like to repeat it three times. This objection applied only to utterances that were not changed in any way. After two repetitions of a single syllable, two repetitions of the same syllable plus another one offered no difficulty. It was, therefore, possible to "build up" to a six syllable utterance according to the following pattern without introducing a falling off in improvement between the first and second repetitions of individual utterances:

1
1
1 2
1 2
1 2 3
1 2 3
1 2 3 4
1 2 3 4
1 2 3 4 5
1 2 3 4 5
1 2 3 4 5 6
1 2 3 4 5 6

(Numbers indicate successive syllables in a six-syllable sentence.)

Variations on this pattern were therefore used in constructing testing and training materials.

The programs were deliberately designed to repeat the same syllables over and over again. The aim was to give maximum opportunity to improve phonological accuracy through self-correction. Only two six-syllable sentences were introduced in each daily Practice Session, with the result that each of the twelve syllables occurred from 16 to 24 times in a Practice Session, depending on its position in the sentence.

Partly to compensate for the monotonous effect of the repetitions (as well as to provide a measure of daily progress) the subjects were given a Pre-test and Post-test before and after each Practice Session and were encouraged to strive to improve their pronunciation so as to do well on the Post-test. They were also encouraged to do their best on both the Pre-test and Post-test. In general, the subjects were highly cooperative, and most of them appeared to be putting forth their best efforts at all times.

Two methods of scoring to measure improvement in phonological accuracy were employed, the Overall and the Phonemic. The former scores were secured by rating the final presentation of the first three syllables of a sentence as well as the final presentation of the entire six-syllable sentence for overall approximation to the French phonological system, including intonation as well as correctness in the production of all the phonemes in the sentence. The Phonemic score was derived from two ratings of a single phoneme, located in the sentence. (See section on scoring, and also Appendix I.) Six target phonemes were selected for scoring, namely /o/, /y/, /ɛ̃/, /ɔ̃/, /l/, /t/. They were chosen because they generally demand the greatest amount of phonetic adjustment on the part of an American speaker. Two of the target phonemes were scored in each of the daily Training Sessions, so that each was tested twice during the training. The training and testing utterances were selected so as to place each target phoneme at one time or another in the initial, medial, and final position.

In planning the short-delay presentations, the question arose as to whether it would be better for the subject to hear the model's voice immediately before the "echo" - that is, the playback, of his mimicry in order to maximize discrimination of differences between them or to hear the echo immediately after his own mimicry so as to mimimize delay of information. The patterns for a single utterance embodying these two alternatives are as follows:

(1) model's voice: S's mimicry: model's voice: S's echo

(2) model's voice: S's mimicry: S's echo

After trying out these patterns, the second was chosen. The following assumptions stood in its favor:

1. It was the simplest pattern and therefore likely to be the least confusing.

2. It was the shortest and therefore permitted practice on the greatest number of utterances in a given period of time.

3. Considerable experimental evidence exists to indicate the advantage of the shortest possible delay between performance and information as to success (or reinforcement). Delay between the model's voice and the echo should not be as serious a handicap as delay between mimicry and echo because the image of correct French pronunciation should be fairly well established by frequent repetitions throughout the period of practice of the French syllables involved. The incorrect aspects of a given echo should be readily discriminated in contrast with a fairly well established image. But to correct a particular error of pronunciation, S would need to compare the sound of the echo with his image of the particular motor pattern that produced the sound. The nearer in time the particular motor pattern to the particular sound, the more effective the comparison might be.

These assumptive considerations appeared to be justified by the experience of the psychologist experimenter in responding to the two patterns. Subjectively, the second pattern seemed easier and less confusing. Errors in pronunciation with this pattern were readily detected in the echo, more clearly than from the immediate feedback occuring during mimicry. Full perception of the errors took an instant of time following the echo. This perception was followed by a set to correct the pronunciation. This set seemed to be "firmed-up" and guided through hearing the model's voice repeat the utterance either as the whole or a part of the next utterance in the series.

The actual mode of arranging for a delay between the end of the model's utterance and the echo to provide for S's intervening mimicry has been described in the section on equipment.

Another problem arose in programing the practice materials so as best to equalize the practice programs for the SD group and the others. Since SD required time for the echo, not as many utterances per unit of time could be programed for it as for the treatments in which only the mimicry was interposed between one model utterance and the next. If longer programs were prepared for the IF, AF and LD treatments, however, a differential factor of program design would be introduced. There seemed to be no way of equalizing the program design for all four treatments unless all four were given the same program. The decision was made, therefore, to adopt the solution that would be most convenient for administering the experiment in the Practice Sessions. The same practice programs were made for all treatments. The model's utterances were widely enough spaced to allow for the echo on SD, and the subjects in the other groups were instructed to mimick each utterance two or three times, whichever seemed best to the subject himself.

Only experimental tests could actually determine the question of whether this arrangement favored the SD treatment or the other three, since repeated mimicry of a single utterance might actually be advantageous. Ideally, the optimum method for each treatment should have been used, but in the absence of knowledge, the adoption of the most convenient solution seemed justified.

To provide for playback on LD, the practice program was repeated each day, and while the other groups practiced it over again, the LD group listened to the recording of its first Practice Session.

During the second day of training in Experiment 3, one of the subjects in the LD group began to repeat the utterances as she listened to the playback. She said she did so to try to correct the mistakes she heard and also to keep her attention from drifting. Although the morale of all groups appeared to be high, definite signs of inattention had been noted in the LD group during playback. To control the factor of attention, all LD subjects were instructed to repeat utterances as they listened to playback, and this instruction continued to be given throughout the series of Main Experiments.

Since the AF condition is obviously considered a method of improving conditions for self-correction, it was decided that the SD and LD groups would work with inactivated headphones, so that each group would be used to test only one method of putative improvement.

Preliminary work with subjects who were completely unfamiliar with the French language showed that they often completely misinterpreted the phonemes of utterances, which, of course, were meaningless to them. A common misinterpretation was "s" for "f" or vice versa. The full English approximations to French vowel phonemes tended to occur, and these were often not the nearest approximations. Once an error had been made, the individual's perception of an utterance seemed to be fixed, and he continued in the error.

This raised the problem of whether our aim should be to test the effectiveness of the four conditions in enabling the individual to correct bad errors and to recognize that French phonemes are different from English phonemes or to test the effectiveness of the varied conditions for achieving a closer approximation to the French phonological system after receiving some phonological instruction. Our conclusion was that in nearly all teaching situations students would be given direct instructions of some sort that would eliminate errors far outside the range of the French phonemes prior to or during practice in the laboratory. We therefore decided to introduce Classroom Session immediately before each Laboratory Session to acquaint the students with the two utterances in the training materials on that day.

Classroom Training

On the first day of each experiment, except Experiment 6, the classroom period was employed to inform the subjects as to the purpose and method of the experiment and to instruct them in various laboratory procedures. They were given the concept of the phoneme and told that no French phoneme is exactly like any American phoneme, hence that they should listen carefully and try to imitate the exact French sounds, rather than translate them into American sounds. This principle was illustrated by the difference between the dipthongized American /o/ and the undiphthongized French /o/ as well as the difference between the French /r/ and the American /r/. They were then briefly introduced to the difference between word and sentence intonation in French and American. These instructions were given to prevent naive Americanization of the French pronunciation in the Aptitude Test so that it might be a test of the student's best mimicry prior to actual training.

For the six Training Sessions, the Classroom Session began with certain general instructions on French pronunciation. At the end of the general instructions, the instructor read through the Pre-test, the build-up preceding the Post-test, and the Post-test, with the class repeating each utterance in unison. (See Appendix I.) No special emphasis was placed on the target phonemes for the day, and the subjects never knew what the target phonemes were. The students were thus acquainted with the practice material for the day under circumstances where they could watch the lips of the model, and gross misinterpretations of the model utterances were thus eliminated. This procedure also made the Pre-tests, as compared with the Aptitude Test, a measure of the improvement effectuated by classroom instruction; whereas the difference between the Pre-tests and the Post-tests served as a measure of the improvement produced by laboratory practice.

The nature of the general instruction given in the first part of the Classroom Session can be illustrated by a detailed account of the procedure on the first training day. The aim was to impress the subjects with the essential differences between American and French speakers in the posture and action of the vocal muscles.

Attention was first called to the absence of gliding or diphthongizing in French vowels. The instructor went through the following series of pairs of English and French words, asking the class to listen carefully to the differences between them.

dear - dire

tea - ti

dough - dos

steel - style

bah - bas

paper - papier

low - lot

Then he had the class mimick him after each of the above words, cautioning them to avoid the glide in the French member of each pair.

Next the instructor called attention to the sharpness or "ping" of the French consonants, pointing out the manner in which the lips are kept firm and crisp. Calling on the class to note both the absence of glide in the vowels and the sharpness of the consonants, he went through the above list again and followed it by having the class mimick him.

Finally, he called attention to the economy of breath and absence of aspiration after consonants and emphasized how the French tongue position in pronouncing /t / and /d / avoids aspiration. Again calling attention to all three of the above features of French pronunciation, he went through the above English-French word pairs as before.

To extend the subjects' practice of the French form of posture and action the instructor finished the general training for the day with:

> "The following French words will be pronounced twice each for further illustration and imitation. As you watch the speaker's lips and tongue carefully, try to imitate the sounds accurately. Especially control your voice and breath very carefully. Ready?"

The following list was read:

dis, tir, de, thé, dammer, tas, dot, tort, dos, tôt, doux, tous, deux, teuton, du, tu.

The instructor then presented the training sentences for the day as previously described.

In the succeeding sessions, subjects were introduced to the whole range of French phonemes. Specific tongue and lip positions for specific phonemes were not taught (except as indicated above). Instead, the general characteristics of French vocal posture and action were re-emphasized, and the subjects were left to learn for themselves through mimicry and self-correction the particular values of each phoneme. Thus, the subjects were given a general approach to learning French pronunciation that was expected to enable them to make progress, but room was left to test the effectiveness of the four treatments to bring about improvement through self-correction, and coaching was not employed to produce specific phonemic accuracy.

In sum, the first part of the Classroom Session was used to establish and re-establish at the beginning of each practice session a general set toward the correct pronunciation of French. The second part was used to introduce the subjects to the training material of the day in a manner that would practically eliminate gross misinterpretation of the French sounds and reduce difficulties in the organization of syllable sequences without affording specific instruction in how to produce either the target phonemes or the others.

During the last three of the six training days, short bits of conversation, the meaning of which was explained, were introduced into the general part of the training session to relieve the monotony of continually mimicking material that was meaningless to the subject. This was not done earlier in order to establish thoroughly the set toward purely phonological mimicry.

Schedule for Daily Training Sessions

The program for each training day for the Replication Experiments was as follows:

1. Classroom Session (about 15 minutes, varied with the needs of the class on a particular day).

2. Laboratory Session (about 30 minutes).
    a. Warm-up and Pre-test (1½ minutes)
    b. Practice Session (9 minutes)
    c. Rest pause (8 minutes)
    d. Repeat of Practice Session with LD listening to the recording of the model (9 minutes).
    e. Warm-up and Post-test (1½ minutes)

(The times are approximate and varied slightly with each day's material.)

The material used in the warm-up before the Pre-test each day was different from the practice material and was used to make sure that the equipment was in working order, the gain settings correct, and the subjects accustomed to the situation before beginning the Pre-test. The warm-up before the Post-test involved build-ups of the two practice sentences and was designed to re-accustom the LD and SD groups to the conditions that the IF and AF groups had been working with to avoid any handicap to them that sudden changes in the form of practice might entail. (See Appendix I.))

The Pre-test was given with all but the AF group in the ordinary IF condition. Between Pre-test and Training Series, the connections to the SD machines were changed to the short delay arrangement. During the rest session, the tapes on the LD machines were wound back to the beginning of the recorded part of the training session and set to play back into the earphones. After the second training session, both SD and LD equipment was set for the ordinary IF condition during the Post-test. These procedures, which were carried out by the proctors, provided a short rest between Pre-test and training sessions and between training sessions and Post-test. The entire procedure for a single day required about an hour unless special delays occurred.

In Experiment 6, the Classroom Session was omitted to test its effect by comparing performance on Experiment 5 with that on Experiment 6.

The differences between the program for Experiment 3 and that for the Replication Experiments will be indicated in the section on Testing and Training Materials.

Testing and Training Materials

The testing and training materials both for the Trial Experiment (Experiment 3) and for the Replication Experiments (Experiments 3, 4, and 7) as well as the Training Program for Experiment 6 are shown in Appendix I. The same Aptitude-Criterion test was employed in Experiment 6 as in the Replication Experiments.

Both test tapes and training tapes were produced by a native French-speaking woman whose style of utterance was exceptionally clear, crisp, and deliberate. Those who have heard these tapes are agreed that they have never heard a model voice that sounded easier to imitate.

In the Main Experiments the Aptitude-Criterion test was administered as an Aptitude Test at the beginning of each experiment before the treatment groups were separated and also as a Final Test after the six days of training. To avoid any bias favoring either inactivated or activated conditions, the first half was administered with activated headphones and, after a rest of eight minutes, the second half with inactivated headphones.

Each half was composed of a warm-up sentence and twelve scored sentences, all of six syllables each. Each of the six target phonemes appeared for scoring in four of the sentences. These sentences were built up one syllable at a time from one syllable to six according to the pattern shown in Appendix I. This style of buildup was selected after trying out various other styles and finding them less well-adapted to the capacities of the subjects.

Every second sentence in the A-C test was used in the daily Training Sessions and was tested with exactly the same kind of buildup in the Pre-tests and Post-tests. Each of the six target phonemes appeared for scoring in two of the twelve sentences of the Training Program.

This procedure resulted in four kinds of tests that served as criteria of learning: (1) the sum of the scores for all six days on the Pre-tests, (2) the sum of all the scores on the Post-tests, (3) the trained utterances on the Final Test and (4) the untrained utterances on the Final Test. (Hereafter referred to as Pre-tests, Post-tests, Final Trained, and Final Untrained, respectively.) The Final Untrained would obviously serve as a measure of the generalization or transfer of learning from the trained sentences to utterances that had not been practiced.

As shown in the Appendix, the first three syllables of an utterance were presented one after another, each being presented twice, at the beginning of the buildup. This was to permit scoring the second occurrence of one of the three syllables for the correctness of a target phoneme. All three syllables were then presented twice, and the second occurrence was scored for the correctness of a single phoneme and for overall phonological correctness. The fourth and fifth syllables were then added one at a time, and then the entire six-syllable utterance was repeated twice and the second occurrence scored for overall phonological accuracy.

To avoid monotony, the first part of each Training Series was devoted to one-syllable utterances, unrelated to the two training sentences, containing the six phonemes on which the week's training was concentrated. This, of course, provided for some degree of generalized training. The variation in types of buildup and the alternation of utterances, which may be seen by inspection of the Training Series, were also aimed at avoiding monotony.

In Experiment 6, as shown *in Appendix I, four sentences* were tested and used in the Training Series each day. Hence, the entire set of sentences in the Aptitude-Criterion test were trained in this experiment. Otherwise, the laboratory procedure in Experiment 6 was the same as in the Replication Experiments. The entire Aptitude-Criterion test was given the first day, followed by six days of training with Pre- and Post-tests, and then the Aptitude-Criterion test was administered the final day. The first day's administration of the Aptitude-Criterion test must be viewed as a criterion, rather than an aptitude measurement. It actually measured degree of retention of pronunciation skills over a week of non-practice. It will, therefore, be referred to as the Introductory test.

The testing and training materials employed in the Main Experiments represented a rather complete revision of those used in Experiment 3. Since the differences can readily be observed by examination of the materials in Appendix I, it will not be necessary to discuss them in detail. Briefly, the target phonemes were different because after scoring the Experiment 3 phonemes, it was decided that the phonemes should be changed to those offering the greatest difficulty. This necessitated a complete change of sentences.

In Experiment 3, the Pre-tests were not announced as such but appeared to the subjects simply as part of the practice. There were two practice sessions, in the second of which the LD subjects heard their first session utterances played back. Then the "Review and Post-test" was administered to all three groups, but the fact that they were being tested was not stressed as much as was later done in the Main Experiments.

In Experiment 3, the principle of confining practice to three-syllable utterances was not employed. The method used to give a maximum opportunity for hearing phonological errors and correcting them was (1) to present single syllables several times (2) to include three-syllable utterances in which there was a short pause between each of the three syllables. This last method actually appeared to increase difficulties. The subjects found it harder to remember the utterances when the model spoke them in this way, and also had trouble with motor coordination, since their natural rate of speech was different from that of the model. The failure of this method led to the decision to use the three-syllable form of training in the Main Experiments.

Scoring

Two variables were scored: (1) the Phonemic variable (Ph), measuring the phonological correctness of single target phonemes and (2) the Overall variable (OA), measuring the phonological correctness of an utterance as a whole.

Two of the scorers (C and S) were experienced French language teachers with considerable special training in phonetics. The third (B) was a native French speaker with three years experience teaching French to American students. C and S scored each utterance independently on the Ph variable. C and B scored the OA variable in similar fashion. All scorers spent considerable time scoring together, comparing results, and arriving at agreement concerning their interpretation of the standards before beginning to score for the Training Experiments.

Each utterance for all three variables was scored by ratings on a seven point scale that was found in the course of preliminary trials to be most convenient for the raters. The standards of scoring were as follows:

3.0: Almost native French
2.5: Between 3.0 and 2.0
2.0: Not correct, but more French than American
1.5: Between 2.0 and 1.0
1.0: Almost wholly American
0.5: Between 1.0 and 0.0
0.0: Badly garbled or wholly American

To avoid halo effects and bias on the part of scorers through knowledge of the treatment or test that was being scored, the records made by the student were transcribed onto special scorers' tapes from a Revere recorder to an Ampex 500 in the following manner:

A random order was established for the utterances of the subjects on the Pre-tests and Post-tests combined and also for the Aptitude Test and Final Test combined. The utterances to be scored were transcribed onto the scorer's tapes in this random order, transcribing only six utterances from two successive sentences at a time. The utterances were identified for the scorers only in terms of their order on the scorers' tapes.

The scorers, therefore, heard and rated the one-syllable, three syllable, and six syllable utterances for a subject for two successive sentences. They then heard and rated the same series of utterances for the next subject. They had no knowledge of the treatment group the subject belonged to or of whether the utterances came from the Pre-tests or the Post-tests in one case or from the Aptitude Test or Final Test in the other.

For each variable, therefore, the scorers rated only four utterances at a time. This was arranged both to prevent halo effect or stereotyped rating and to permit the scorers to concentrate on a few utterances at a time. C, who scored both the Ph and OA variables, first went through each tape for the Ph variable and then played it through again for the OA variable.

In order to achieve comparable scoring throughout Experiments 4 through 7, it would have been preferable to arrange the utterances for all four experiments and all four tests in random order, but considerations of time made this impracticable, since it was necessary for the scorers to rate the utterances on one experiment while the later experiments were in progress.

After the scorers had completed their ratings, these ratings were punched on tabulating cards, with the ratings for one utterance only on each card, according to the following scale:

Scorer's Rating	Number on Card
3.0	7
2.5	6
2.0	5
1.5	4
1.0	3
0.5	2
0.0	1

The score for a subject for each variable on each item in the test was secured by summing the card numbers for both scorers. Thus the score on an item could vary from 2 to 14.

The Questionnaire

Immediately after the Final Test in the four Main Experiments, a questionnaire, shown in Appendix II, was administered. It was designed to determine the general state of morale and motivation throughout the experiment as well as to find what features of the procedure had tended to depress morale and what features had tended to improve it.

The subjects were asked not to put their names on the questionnaire, and they were encouraged to be as frank and objective as possible. To overcome the tendency toward kindness or politeness in responding to questionnaires, it was emphasized that critical comments would be of genuine value to the experimenters in enabling them to correct their errors of procedure and that the only way in which such errors could be determined was through such a questionnaire as this.

VI. RESULTS AND DISCUSSIONS

Since the entire investigation involved a considerable number of complex procedures, it seems advisable to follow the report of results on each major part of the investigation with a discussion on that part. This section will begin, therefore, with the results of the questionnaire, which will throw some light on certain outcomes of the experimentation. A report on the Replication Experiments, which constituted the heart of the investigation, will follow. The Continuation Experiment and Trial Experiment will then be reported and these reports will be followed by a general discussion.

The Questionnaire

Results

Following an item which identified the treatment group, the questionnaire contained four choice response items which may be identified as follows:

Item 2:	Estimate of educational value
Item 3:	Rating of interestingness
Item 4:	Rating of boresomeness
Item 5:	Rating of effort

These items were scored on the following scale running from high to low indices of motivation or morale:

Score	Item			
2	2c	3a	4a	5a
1	2b	3b	4b	5b
0	2a	3c	4c	5c
-1				5d

Mean scores per item are shown in Table II. To make their interpretation more meaningful, the following choices on the questionnaire are given together with their scores:

Item 2: What I learned in this experiment
- c. Will definitely be of value to me (Score,2)
- b. May be of some value to me (Score,1)

Item 3:
- a. All the work I did in this experiment was interesting (Score,2)
- b. Some of the work I did in this experiment was interesting (Score,1)

TABLE III

MEAN SCORES PER ITEM WITH CROSS MEANS FOR EXPERIMENT, TREATMENT AND ITEM ON THE CHOICE RESPONSE ITEMS OF THE QUESTIONNAIRE WITH LISTING OF SIGNIFICANT VARIANCES.

Experiment X Treatment

	IF	AF	LD	SD
Ex 4	1.5	1.4	1.6	1.5
Ex 5	1.3	1.3	1.1	1.3
Ex 6	1.0	1.3	1.1	1.3
Ex 7	1.5	1.6	1.6	1.7
Mean	1.3	1.4	1.4	1.5

Experiment X Item

	It2	It3	It4	It5	Mean
Ex 4	1.3	1.7	1.2	1.8	1.5
Ex 5	1.1	1.2	1.0	1.8	1.3
Ex 6	1.1	1.0	1.0	1.6	1.2
Ex 7	1.2	1.9	1.4	1.9	1.6
Mean	1.2	1.4	1.2	1.8	

Treatment X Item

	It2	It3	It4	It5	Mean
IF	1.2	1.4	.9	1.7	1.3
AF	1.1	1.5	1.2	1.8	1.4
LD	1.2	1.3	1.2	1.7	1.4
SD	1.2	1.4	1.3	1.9	1.5
Mean	1.2	1.4	1.2	1.8	

Significance of Variances

Treatments	N.S.
Experiments	$P < .001$
Items	$P < .001$
I x E	$P < .001$
I x E x T	$P < .001$

Item 4: a. None of the work was boring (Score,2)
b. Some of the work was boring (Score,1)

Item 5: a. I did my best almost all of the time (Score,2)
b. I did my best more than half the time (Score,1)

The means of the various responses to the items for the four groups combined show that most of the subjects claim to have done their best most of the time, about half say that all of the work was interesting and about half that some of the work was interesting. The responses cluster toward the statement that what was learned "may be of some value to me", with a stronger tendency to select "will definitely be of value" than "will never be of any value." Similarly, the responses approximate "some of the work was boring", with a greater tendency to say "none of the work was boring" than "some of the work was very boring."

There is a significant difference between experimental groups on the total score for the questionnaire, with the high school groups indicating a higher level of morale than the college group. About half of this difference is accounted for by the greater degree of interest expressed by the high school groups. The average response for the college group is "some of the work was interesting." The average for the high school groups approaches "all of the work was interesting." On the other hand, the college group lays claim to about as much effort as the high school groups. These differences in response to the items account for the statistically significant interaction between items and experiments.

Examination of individual replies and the frequency of responses between individual items and treatment groups in separate experiments suggest that the cause of the statistically significant interaction between items, experiment, and treatment is the fact that a few subjects showed a different pattern of response to the items than was characteristic of their experimental group. As might be expected, there were real differences between individuals in the way they ranked the various items.

No significant difference was found between treatments. There is a non-significant trend for the IF groups to find the experiment more boring and for the SD group to be generally higher than the others.

To analyze the open-ended questions, the responses were placed in what appeared to be the most meaningful categories. No categories were included that did not contain at least five responses in all four groups, and the remaining responses were categorized as miscellaneous. Appendix III shows the categories, together with the number of responses falling in each category by experiment and treatment.

The responses shown in Appendix III were categorized into "high morale" responses and "low morale" responses, with the responses to the question "What bothered you so that you couldn't do your best work?" omitted because it was difficult to determine whether these responses stemmed from high motivation or low morale. The proportions of high morale responses for experiments and treatments are as follows:

Ex 4 --- .621   IF --- .567
Ex 5 --- .429   AF --- .655
Ex 6 --- .552   LD --- .521
Ex 7 --- .732   SD --- .600

The higher morale of the high school as compared with the college subjects is confirmed by this analysis. The fact that the order of morale for the treatment groups changes from that secured by analysis of the choice response questions points to a lack of any real difference in morale between treatments.

Discussion

The finding of no significant morale differences between treatments is important since it eliminates the factor of morale or motivation in explaining any experimental differences in efficiency between treatments.

In discussing the differences between experimental groups, the item number and category of response letter in Appendix III will be referred to as follows: (1A) would refer to the first response category of the first item, namely "Testing (usually pre-post-tests) or observing own progress." This item, the most frequently mentioned point of interest, points to the high level of achievement motivation which seems to have characterized all groups, and which is also indicated by the high scores on the fifth item of the choice response section.

The repetitiousness of the practice sessions, which was deliberately introduced for purposes of stressing training in phonological accuracy, appears to have been more distressing to the college group than to the high school groups, particularly the junior high school group, although the repetitious practice sessions were boring to some individuals in all groups (2A,2B,2D, 3D,4D,5D). The rather elaborate checks to make certain the machines were functioning properly at all times, together with the constantly repeated instructions (the former found necessary during the trial experiment and the latter adapted to the junior high school level of requirement) were irritating to some of the college group, but not at all to the junior high school group (2E,3B,5F).

After observing the three groups, the experimenters have come to the conclusion that the process of mimicry itself is intrinsically uninteresting to most students of college level, whereas it was intrinsically interesting to the junior high school group. The senior high school group seemed to stand in an intermediate position in this respect. Evidence for these conclusions is found in IE,IF,2C,3A, as well as in the lower degree to which the high school groups complained of repetitiousness. The junior high school group appeared to be genuinely challenged by the problem of pronunciation. They felt it to be difficult, but worth struggling with (3F,4C). The senior high school group was not satisfied, as the junior group was, with mere mimicry. They wanted to learn the language, hence their greater appreciation of the classroom sessions, and their objection to not seeing the written language or knowing the meaning of the sentences (1C,3E,5B,6D,6J). Throughout the experiment, the senior high school group begged to know the meanings of the sentences, and they showed great satisfaction in learning them at the end in the course of a small party that was given as a reward for their good work. The junior high school group paid little attention to the cards on which the meanings of the sentences were given, but throughout their party continued to babble the sounds that they had been practicing.

It seems possible that the better progress that young children make in learning to pronounce a language is partly based on a greater intrinsic interest in the mere production of new sounds. Probably learning to pronounce through mimicry of a foreign voice is a more meaningful and interesting procedure the younger the student.

For both groups of high school students, the whole experience was something of a pleasant adventure. Coming to a speech laboratory in a university was an exciting novelty, especially for the senior high school students who thought of college attendance as the next great step in their course of growing up. The fact that the staff in this prestigeful place showed a friendly interest in them was genuinely pleasing, and this may in part have accounted for their special liking for the classroom sessions (1C,6C,6D). To both high school groups, the experience had many of the social as well as experiential values of a field trip. During the rest pauses, they bought soft drinks at the vending machine and had a good time generally; whereas the rest pauses for some of the college students were periods of boredom and impatience to get on with the work (3G,6B).

There was little glamor in the experience for the college students. They were typically very busy young men who were not getting enough sleep (4E), and they were eager to get the job out of the way as expeditiously as possible (3G,5F). Most of them did not plan ever to study French. They were uninterested in what they were learning, but many of them seemed to feel a genuine interest in the scientific and practical value of the work. Hence, almost the only thing they found to like about the experiment was its efficiency and good planning (6A), which it occurred to fully half of them to mention. There was every evidence that most of them were interested in doing a similar good, efficient job out of a feeling of both pride and obligation. Learning, however, probably goes on more effectively when the task is pleasant, and the college students may have been handicapped in learning by an absence of pleasurable feeling in spite of their willingness to invest effort.

In closing, three minor outcomes of the questionnaire may be mentioned. The senior high school group was naturally somewhat disturbed by the occasional malfunctioning of the SD machines (3C,4F). The senior high school group more often mentioned outside disturbance or self-consciousness about being heard as being bothersome (4B). Finally, although the playback groups did not have substantially higher morale, some students mentioned hearing their own voices on playback as being interesting or helpful (1B,6H).

The Replication Experiments (Experiments 4, 5 and 7)

Results

In reporting the results of all experiments, the unit used will be the per item score from the tabulating cards multiplied by 100 to avoid decimals. Thus the meaning of all averages for the Phonemic and Overall variables can be understood in terms of the standards described in the section on scoring. As indicated on page 31, the scores could run from 2 to 14, a score of 2 being equivalent to a rating of 0.0 by both raters and a score of 14 to a rating of 6.0 by both raters. According to the scoring standards described on page 30, scores on the per item times 100 scale have the following meaning:

1400	Almost native French
1000	Not correct, but more French than American
600	Almost wholly American
200	Badly garbled or wholly American

Nearly all the averages that will be reported in this section lie between 600 and 1000 and represent some degree of progress between "almost wholly American" to "Not correct, but more French than American". This, of course, is what might be expected for a group of beginners.

It will be useful to have some standard for judging the amount of real difference a given score difference between tests or treatment groups represents. The most meaningful unit of this sort is the standard deviation. It was found that the mean standard deviation of an experimental group for all experiments and all tests was 77 in terms of the per item times 100 scale and that there was no appreciable difference between the mean SD's of the aptitude tests and the criterion tests or between the mean SD's of the Phonemic and the Overall variables. It was therefore decided to use 80 as a "standard unit" in estimating the importance of all changes from test to test and all differences between treatment or experiment groups. This unit of 80 per item times 100 points represents one fifth of the distance between "Almost wholly American" and "Not correct, but more French than American." To facilitate thinking in terms of this unit, indices in all charts are shown in simple fractions of the unit.

Table IV shows a series of abbreviations frequently used in reporting means.

The bulk of the results of statistical analysis for this section has been placed in Appendix IV. Table A shows a series of reliability coefficients computed for the Aptitude Test. The split-halves were secured by correlating odd-even items, the odd items were those used later in the training series. The two

TABLE IV

FREQUENT ABBREVIATIONS USED IN TABLES AND CHARTS.

(All scores shown in the tables are mean scores or adjusted mean scores. Units are always per item x 100 as described in the text.)

Ap:	Aptitude Test (Aptitude-Criterion test administered at the beginning of the experiments)
Pr:	Pre-tests (All six Pre-tests administered at the beginning of each laboratory training session)
Po:	Post-tests (All six Post-tests administered at the end of each laboratory training session)
FT:	Final Trained (Sum of items in Aptitude-Criterion test administered at end of experiment which were used in training series.)
FU:	Final Untrained (Sum of items in Aptitude-Criterion test administered at end of each experiment which were not used in training series.)
IF:	Inactivated Feedback (Group with Inactivated Feedback treatment)
AF:	Activated Feedback (Group with Activated Feedback treatment)
LD:	Long Delay (Group with Long Delay treatment)
SD:	Short Delay (Group with Short Delay treatment)
Ex4:	Experiment 4 (Group in Experiment 4)
Ex5	Experiment 5 (Group in Experiment 5)
Ex7	Experiment 7 (Group in Experiment 7)
Ph:	Phonemic variable
OA	Overall variable
Co:	Combined variable (Overall and Phonemic variables added together)
M:	Mean (Mean scores for any group)
Mm:	Mean of Means (Mean of any row or column of means)
Dm:	Deviation from Mean of Means (Deviation of a given mean from the mean of a set of means)
MDm:	Mean Deviation from Mean of Means (Mean of the deviations of a set of means from the mean of the set)

halves had been made closely similar by including exactly two of the target phonemes in each half.

Examination of this table shows a satisfactory level of reliability for both raters combined throughout all three experiments, although there is some decrease in reliability in the later experiments, especially for the Phonemic variable, probably resulting from rater fatigue. The Overall ratings are consistently more reliable than the Phonemic and are as reliable as the Phonemic and Overall ratings combined.

Except for the Phonemic ratings in Experiment 5, rater C is consistently more reliable than the other two. However, the correlations between raters, especially after correction for attenuation produced by the imperfect reliability of their ratings, shows that the raters were rating on the basis of the same concept of good pronunciation.

Nevertheless, although the raters agreed with one another closely as to the relative standings of the subjects, they interpreted the instructions differently with respect to the anchorages of their ratings.[1] As shown in Table B, Rater S rated consistently below Rater C and Rater B rated consistently above Rater C. Rater B's anchorages definitely shifted upward relative to C's from experiment to experiment. The differences are far above the random fluctuations found for the self-differences of the raters on the odd and even items of the Aptitude Test.

Because of these differences in anchorages, no interpretation can be given to average differences between the two variables; and because of shifts in anchorages from experiment to experiment, average differences between experimental groups are uninterpretable. This does not, however, interfere with comparison of differences between the groups in amount of improvement over the Aptitude Test shown in their criterion tests.

Table C in Appendix IV shows the intercorrelations between the Phonemic and Overall variables for the various tests. The corrections by the Spearman-Brown formula for the criterion tests were made because these tests contained only half as many items as the Aptitude Test. Even without these corrections, there is a definite tendency for the two variables to be more closely associated with one another in the Post-test and Final Trained than in the Aptitude Test, except in Experiment 5.

Table D, Appendix IV, shows intercorrelations from one test to another. The lower part of this table shows certain means of these correlations which were computed by the z-transformation method to bring out the average commonalities for three kinds of tests, as follows:

---

[1] See the definition of "anchorages" on p. 57.

1. Aptitude Test: Tests skill in pronouncing sentences before any instruction or practice.

2. Training Tests; that is, the Pre-tests, Post-tests and Final Trained: Test skill in pronouncing sentences after laboratory practice and/or classroom instruction have improved their pronunciation.

3. Final Untrained: Tests skill in pronouncing utterances other than those used in training subsequent to instruction and practice on the training utterances.

The mean correlations summarize the commonalities of the Aptitude Test with the Training Tests and the Final Untrained and of the Training Tests with each other and with the Final Untrained. Table D shows the mean of all the correlations between Aptitude and Training Tests is .64; between Aptitude and Final Untrained, .73; among the Training Tests .76; between Training Tests and Final Untrained .73. The differences between the Aptitude-Training Test means and the other means are statistically significant. The result indicates that the Training Tests were measuring some factor not correlated with the Aptitude Tests and that the Final Untrained was only partially affected by this factor.

Interestingly, the Phonemic Aptitude seems to have been a better predictor of later scores than the Overall Aptitude; whereas the Overall Training tests were better predictors than the Phonemic Training tests.

The examination of Tables C and D led to the conclusion that the relationship between the variables was sufficiently complex to justify an analysis of experimental differences for both variables separately. In the outcome, however, both variables produced essentially similar experimental results. The data on the separate analyses, therefore, have been placed in Appendix IV, and the tables shown in the text are chiefly comparisons of combined means of the two variables or means derived from combined scores.

Table E, Appendix IV, shows the raw means for both variables for all treatment groups in all three experiments and also the differences between the Aptitude means and the means of the four criterion tests. These data are condensed in Table V and Fig. 4 which show the means and differences for the Phonemic and Overall variables and the Combined means for the entire population of the three experiments. The only change from test to test for the Phonemic and Overall variables that is not statistically significant at the .01 level of confidence is the slight gain between the Post-test and Final Trained for the Overall variable. In terms of the standard unit of 80 points there is

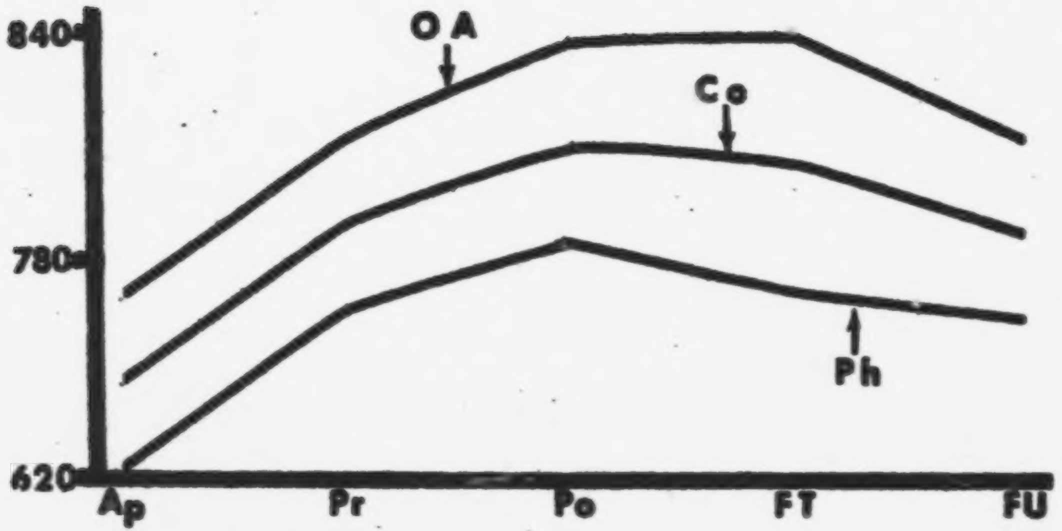

Fig. 4. Mean raw scores, all Replication Experiments combined, for the Phonemic, Overall and Combined variables on the Aptitude Test, Pre-tests, Post-tests, Final Trained and Final Untrained.

TABLE V

MEANS OF PHONEMIC AND OVERALL VARIABLES AND COMBINED MEANS SUMMING ALL THREE REPLICATION EXPERIMENTS WITH DIFFERENCES BETWEEN APTITUDE AND CRITERION TEST SCORES.

Special abbreviations: DPr, DPo, DFT, DFU; differences between Aptitude test and criterion test scores.

	N	Ap	Pr	Po	FT	FU	DPr	DPo	DFT	DFU
Ph	83	632	742	788	757	734	110	156	125	102
OA	83	752	866	932	939	862	114	180	187	110
Co	83	692	804	860	848	798	112	168	156	106

a gain on the Combined variable of 1.39 units between Aptitude and Pre-test, 0.71 units between Pre-test and Post-test, 2.10 units between Aptitude and Post-test, 1.95 units between Aptitude and Final Trained and 1.33 units between Aptitude and Final Untrained.

To determine whether the large gain from Aptitude to Pre-tests should be attributed to the Classroom Sessions which preceded the Pre-tests or to the cumulative effects of six days practice, a computation was made for each Pre-test day separately to determine the gain in mean score over the mean score made on the same items in the Aptitude Test. The results are shown in Table F, Appendix IV. There is a definite tendency for the higher gains on the Phonemic variable to have occurred in the earlier days of training and a slighter tendency on the Overall variable for the greater gains to have occurred in the later days of training. The Combined means show greater gains in the earlier days. These day-to-day differences are probably not statistically significant, but it appears certain that cumulative improvement in the training period does not account for the marked gain on the Pre-tests over the Aptitude Test.

Table G, Appendix IV, shows the means for the four treatment conditions for each of the treatments as they were adjusted in the course of analyses of covariance for each criterion test in each experiment, using the Aptitude Test as the predictor. The treatment variance was found to be significant at the .05 level for only one of these analyses; namely, the Post-test for Experiment 7. In a set of 24 analyses, such a result might be expected to occur as a random variation. Furthermore, the rank order of the treatments varies from experiment to experiment and treatment to treatment. This variation is greater between experiments for a given test than between tests within experiments, suggesting that random errors of group selection for the various treatments may have contributed considerably to the error variance.

Nevertheless, examination of the table reveals a definite average superiority for the Activated Feedback condition and average inferiority for the Short Delay condition. The average rank order for AF is 1.55, for IF 2.25, for LD 2.35, and for SD 2.96.

Table VI shows the general trends in the data of Table G, Appendix IV, secured by combining, that is, taking the means of the Phonemic and Overall adjusted means. The mean deviations are indices of the amount of differentiation among the treatment groups produced either by random errors in group selection or by real differences in the effects of the treatments. As summarized in the right-hand column, where the means of all three experiments are shown, these differences are greatest for the Post-test, next for the Pre-test, then the Final Trained and the Final Untrained. The mean standings of the experimental groups are the same for the first three tests, namely AF first, LD second, IF third, and SD fourth. These standings change in the Final Untrained, with IF first, AF second, LD third, and SD fourth. In short, the first three tests, that is, the Training tests seem to reveal on the average a certain definite pattern of differentiation among the treatments, which tends to disappear in the Final Untrained.

The regularity of these average standings from test to test is displayed in Fig. 5, which diagrams the data shown in the right hand means column of Table VI.

The rank order of the treatments, however, varies from experiment to experiment, as indicated in Fig. 6, which shows the

TABLE VI

COMBINED MEANS OF ADJUSTED PHONEMIC AND OVERALL MEANS OF TREATMENTS DERIVED FROM ONE-WAY ANALYSIS OF COVARIANCE OF EACH CRITERION TEST IN EACH EXPERIMENT IN THE REPLICATION EXPERIMENTS WITH MEANS OF TREATMENT MEANS FOR ALL THREE EXPERIMENTS AND MEANS OF TREATMENT MEANS FOR EACH TEST IN EACH EXPERIMENT TOGETHER WITH DEVIATIONS AND MEAN DEVIATIONS OF TREATMENT MEANS FROM MEANS OF TESTS.

		Ex4		Ex5		Ex7			
		M	Dm	M	Dm	M	Dm	Mm	Dm
Pre-tests	IF	812	+14	790	-39	774	-7	792	-11
	AF	833	+35	862	+33	800	+19	832	+29
	LD	781	-17	838	+ 9	794	+13	804	+ 1
	SD	766	-32	826	- 3	755	-26	782	-21
	Mm	798		829		781		803	
	MDm		24		21		16		15
Post-tests	IF	869	+15	847	-43	844	+ 8	853	- 7
	AF	897	+43	924	+34	854	+18	892	+32
	LD	828	-26	901	+11	859	+23	863	+ 3
	SD	822	-32	888	- 2	788	-48	833	-27
	Mm	854		890		836		860	
	MDm		29		22		24		17
Final Trained	IF	864	+ 8	817	-36	836	0	839	- 9
	AF	897	+41	873	+20	830	- 6	867	+19
	LD	844	-12	857	+ 4	857	+21	853	+ 5
	SD	819	-37	865	+12	820	-16	835	-13
	Mm	856		853		836		848	
	MDm		24		18		11		11
Final Untrained	IF	809	+18	827	+ 7	783	- 2	806	+ 7
	AF	805	+14	832	+12	779	- 6	805	+ 6
	LD	777	-14	822	+ 2	801	+16	800	+ 1
	SD	773	-18	798	-22	777	- 8	783	-16
	Mm	791		820		785		799	
	MDm		16		11		8		7

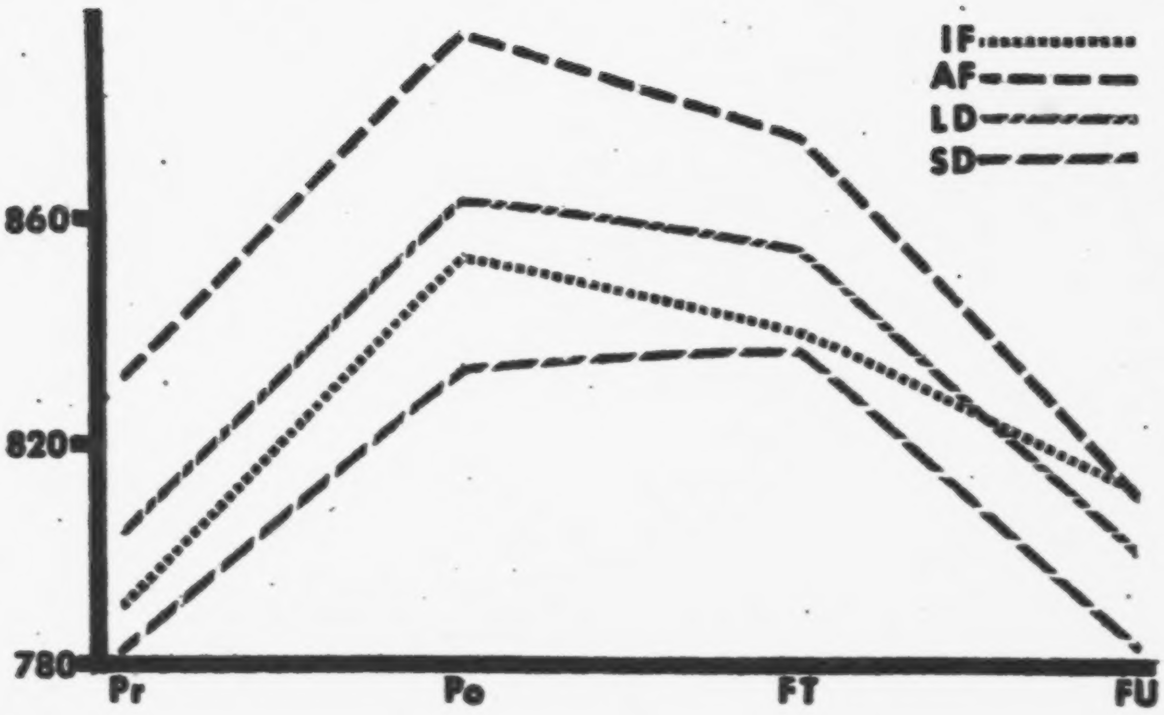

Fig. 5. Adjusted means of treatments, combining all three Replication Experiments and Phonemic and Overall variables for Pre-tests, Post-tests, Final Trained and Final Untrained. (From right hand column, Table VI.)

Combined means of each treatment in each experiment for each test separately. A definite pattern of variation is found between the three experiments throughout the first three tests. This pattern disappears in the Final Untrained.

In all the first three criterion tests IF does relatively best in Experiment 4, slightly less well in Experiment 7, and very poorly in Experiment 5. AF does best in Experiment 4, next best in Experiment 5, and the poorest AF performance is in Experiment 7. LD's relative positions are the reverse of AF, poorest in Experiment 4, best in Experiment 7. SD is very low in Experiments 4 and 7, but hovers around the average of the three groups in Experiment 5. The pattern for each group is more marked in the Post-test than in the Pre-test. In the Final Trained, the pattern shows less differentiation from experiment to experiment for IF, LD, and SD, but the greatest differentiation of all for AF. In the Final Untrained AF and LD retain their patterns, but the patterns for IF and SD disappear. The possible meaning of this peculiar system of regularities from test to test will be dealt with in the discussion section.

As shown in Table VI and Fig. 5, the differentiation between treatments is greater for the Pre-tests than for either of the Final Tests and nearly as great as for the Post-tests. Since the Pre-tests preceded experimentally differentiated

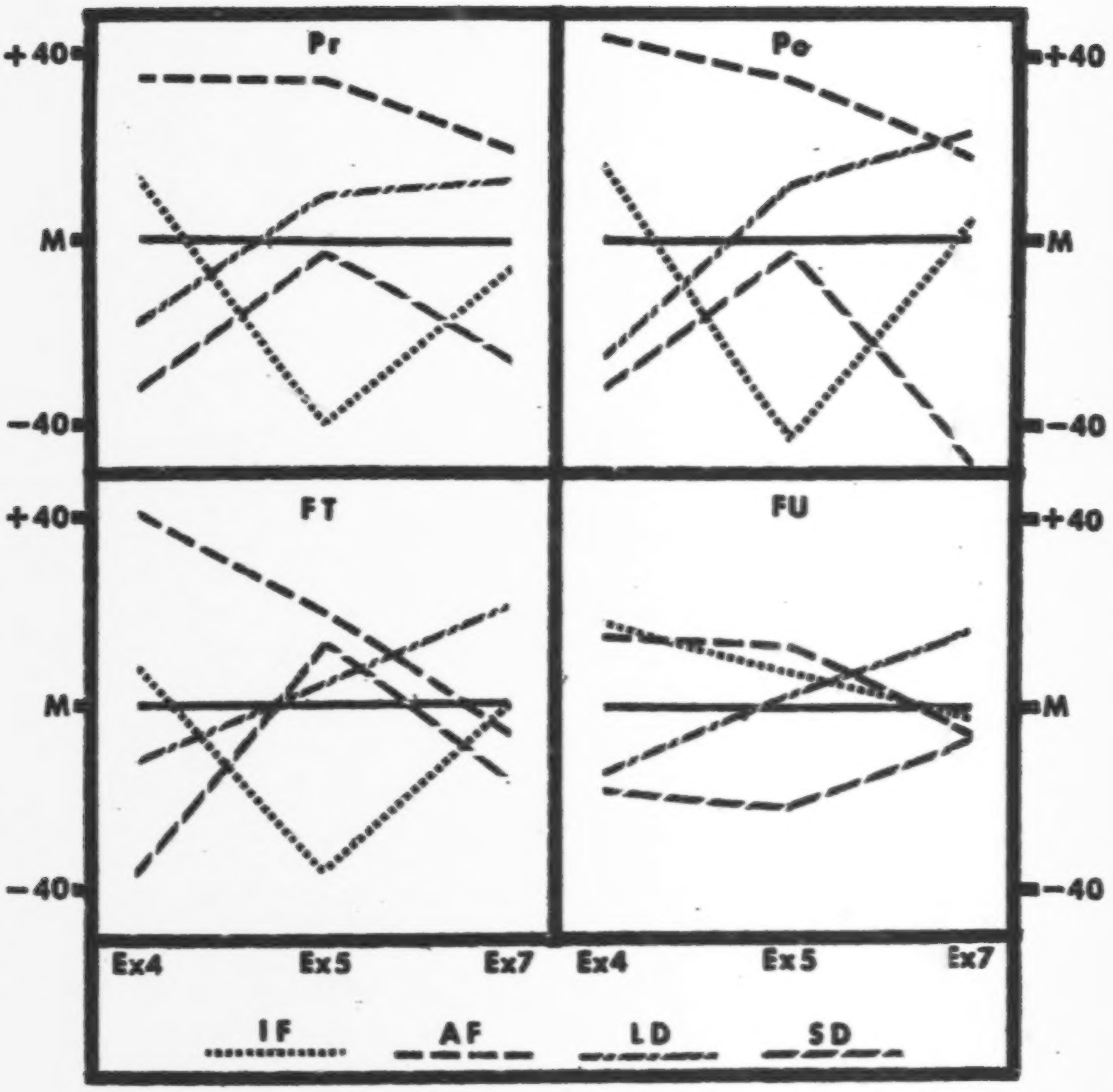

Fig. 6. Adjusted means of treatments, combining Phonemic and Overall variables, showing changes in treatment standings from experiment to experiment for all four criterion tests. (From Table VI.)

practice on each day's work, the sources of this variability could be of only three kinds: (1) random group variation (2) differences in conditions under which the groups took the Pre-tests (3) generalization to the latter days of Pre-testing of differential effects resulting from the earlier days. If this third possibility were true, the treatments would be relatively alike during the earlier days of Pre-testing and would show great differentiation on the latter days. To test the third possibility, mean scores for each of the four treatments on the first three Pre-tests were compared with mean scores on the last three, as shown in Table H, Appendix IV. The results completely disconfirm the third supposition, since the treatment

TABLE VII

ADJUSTED COMBINED SCORE MEANS OF TREATMENTS AND EXPERIMENTS FOR EACH CRITERION TEST OF THE REPLICATION EXPERIMENTS DERIVED FROM TWO-WAY ANALYSIS OF COVARIANCE WITH MEANS OF MEANS AND DEVIATIONS FROM MEANS OF MEANS.

		Pr		Po		FT		FU	
		M	Dm	M	Dm	M	Dm	M	Dm
Treat-ments	IF	789	-18	848	-18	834	-15	801	+1
	AF	829	+22	890	+24	867	+20	807	+7
	LD	802	- 5	857	- 6	850	0	792	-8
	Mm	807		866		850		800	
Experi-ments	Ex4	808	+ 1	858	- 8	868	+18	798	-2
	Ex5	798	- 9	857	- 9	815	-35	784	-16
	Ex7	814	+ 7	884	+18	868	+18	818	+18
	Mm	807		866		850		800	

most superior on the Pre-tests, the AF, is the only one on which there is a combined lower average on the last three days than on the first three.

The combining of adjusted means in Table VI is, of course, a statistically imprecise procedure. It was undertaken to bring out in condensed approximate form the relationships revealed through the adjusted means in Table G, Appendix IV, as a guide to further, more precise analysis. On the basis of this exploratory analysis, it was decided to drop the SD treatment from further analyses, since its inferiority might be explained in terms of the inferior sound quality of the SD machines as noted in the section on equipment. Inclusion of the SD treatment in further analyses might introduce spurious findings of statistical significance based on an experimental error rather than a true experimental effect.

Two-way analyses of covariance were made for the Phonemic variable, the Overall variable and the Overall and Phonemic combined. In the latter analyses the OA and Ph scores were summed for the criterion tests, whereas they were used separately to produce a multiple regression prediction with the Aptitude Test.

The results of these analyses are shown in Table I, Appendix IV, for the Phonemic and Overall variables and in Table K for the combined variable. The results of all analyses are virtually similar as far as treatment variances are concerned. For the first three tests, with the exception of the Final Trained for the Phonemic variable, the treatment variances are larger than the interaction variances. These interaction variances are produced by the variation of the treatment standings from experiment to experiment as shown in Fig. 6.

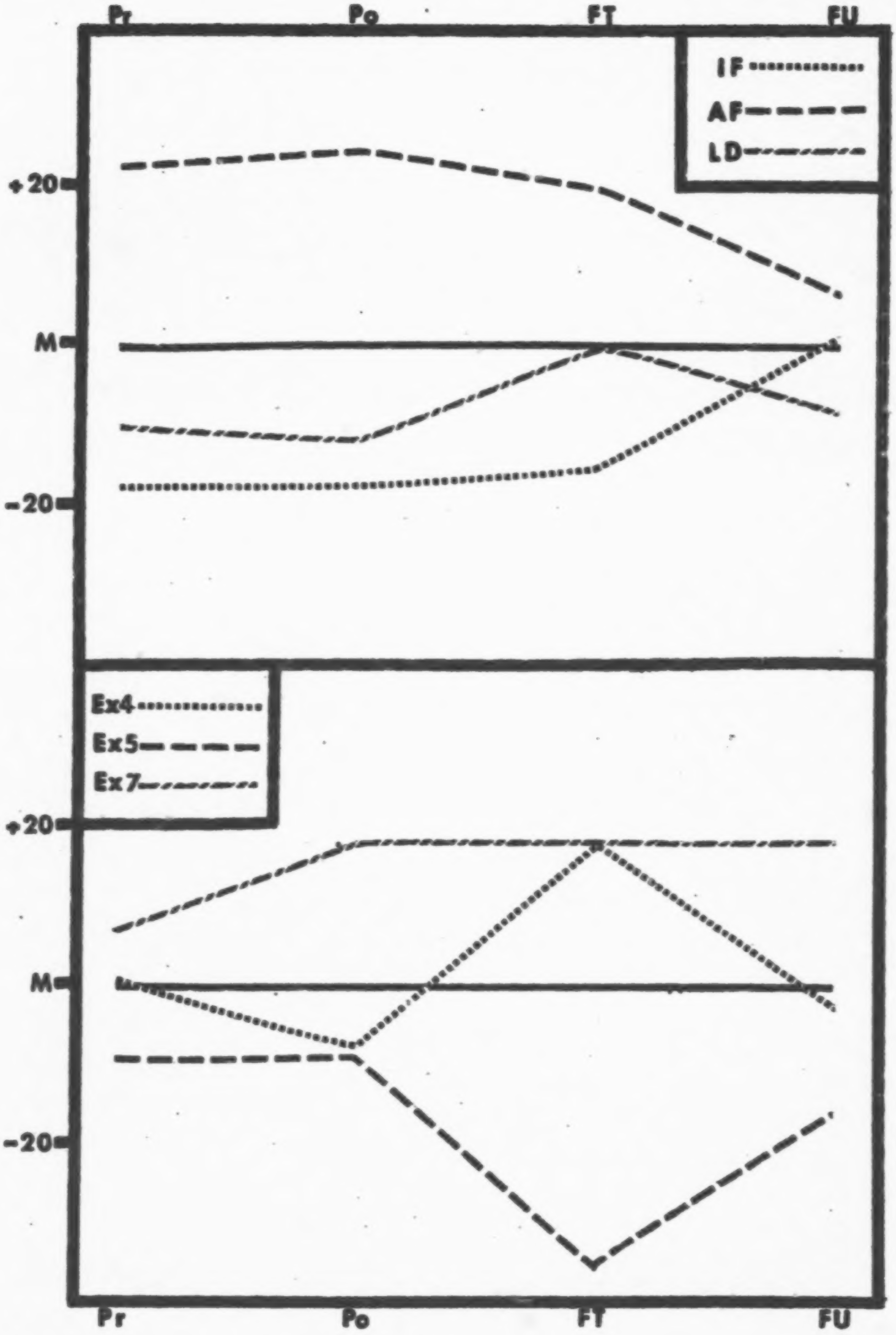

Fig. 7. Deviations from the mean of all three of adjusted means for the Combined variable. Top: Deviations of treatment means. Bottom: Deviations of means of experiment groups. (From Table VII.)

Hence it appears that throughout all three experiments experiment-to-experiment fluctuations in the standing of the treatments are less significant than their average differences throughout the experiments. The absence of substantial treatment variance in the Final Trained on the Phonemic suggests that whatever effect produced these differences generalized rather weakly to the Phonemic variable in the Final Trained. This failure to generalize may be related to the marked decrement in Phonemic scores from the Post-test to the Final Trained shown in Fig. 4.

The analyses show a marginal degree of statistical significance for treatment differences in the Pre-tests on the Phonemic and in the Post-tests on the Overall and Combined variables. The data fulfill the assumption of homogeneous regression which underlies the analysis of covariance, but the Hartley test for homogeneity of adjusted variances in the sub-cells showed significant differences. This casts some further uncertainty on the finding of significant differences between treatments.

Table J, Appendix IV, shows the distribution of adjusted means derived from the analyses reported in Table I. Table VII shows the distribution of adjusted means for the Combined variable, together with the deviations from the mean of all three. Fig. 7 shows the changes in these deviations from test to test.

The consistency of results for treatment variations on the Pre-tests, Post-tests and Final Untrained can hardly be interpreted in any other way than to assume that the differences are produced by the same factor or factors, either by experimentally produced effects or effects produced by random group selection. The somewhat marginal determinations of statistical significance favor the interpretation that the effect was experimentally produced.

For the Combined variable, the range of differences between treatments in "standard units" of 80 is .50, .53, .41 and .19 for the Pre-tests, Post-tests, Final Trained and Final Untrained respectively. The difference between the IF and the AF groups on the Post-tests is one-fourth of the mean gain of 168 for all groups, including SD, between Aptitude and Post-tests as shown in Table V. For the Pre-tests, this difference is 36 percent of the mean gain.

The results reported in Table VIII were secured in order to determine whether the superiority displayed by the AF condition on the Final Test might be attributable to an enhanced capacity for dealing with the first half of the test, which was administered with activated headphones. The table shows that this was the case. If the Final Test had been given entirely in the Inactivated condition, the AF group would have done no better than the others on the Final Trained and would have been somewhat inferior on the Final Untrained.

TABLE VIII

MEAN COMBINED OVERALL AND PHONEMIC GAINS OVER THE APTITUDE TEST FOR THE ACTIVATED AND INACTIVATED CONDITIONS IN THE FINAL TRAINED AND FINAL UNTRAINED TEST FOR ALL THREE REPLICATION EXPERIMENTS COMBINED COMPARING THE ACTIVATED TREATMENT WITH THE INACTIVATED AND LONG DELAY TREATMENTS COMBINED, WITH DIFFERENCES BETWEEN THESE MEAN GAINS.

	Final Trained		Final Untrained	
	Activated	Inactivated	Activated	Inactivated
AF	211	153	132	105
IF + LD	158	146	103	117
Diff.	53	7	29	-12

Note: Aptitude score for AF is 679, for IF and LD combined, 693. Scores adjusted for regression would decrease the raw gain advantage of AF.

In the analyses of covariance, (Tables I and K, Appendix IV) significant differences between experiment groups; that is, college students, senior high school students, and junior high school students are found only for the Final Trained test. As shown in Table J and also Table VIII and Fig. 7, the college group is considerably below the two high school groups. For the Combined variable, this difference is .65 of a "standard unit" of 80. It is slightly over one third of the mean gain from Aptitude to Final Trained for all groups as shown in Table IV.

Discussion

The basic objective of the Replication Experiments, as stated in the section on the analysis of the problem, was to determine the relative effectiveness of the four treatments for learning to pronounce. The only one of the four criterion tests on which a finding of differences between treatments could have shown a genuine superiority of one treatment condition over another was the Final Untrained. The other tests could show only the effectiveness of the treatments in improving the pronunciation of specifically practiced sentences, measuring the improvement in terms of scores adjusted on the basis of the Aptitude Test.

Specifically, the Pre-tests were a measure of the improvement produced by a Classroom Session in which the subjects were instructed in the vocal postures for French pronunciation and in which they heard and saw the instructor pronounce the training sentences and imitated his pronunciation. The failure of the Pre-tests to show greater improvement over the Aptitude Test in the later days of instruction than in the earlier days (Table F, Appendix IV) indicates that practically all the improvement on the Pre-tests was derived from this Classroom Instruction. The Pre-tests were, therefore, measures of *improvement from instruction*. The Post-tests were measures of immediate *improvement from practice*. The Final Trained was a measure of *retained improvemen* on the specific utterances that had been practiced in the previous six days. Only the Final Untrained was a measure of *generalized learning* to pronounce.

It should be noted that there was no experimental differentiation of treatment prior to the Pre-tests. All treatment groups had exactly the same experience in the Classroom Sessions. Hence, the differences found in the scores of the treatment groups on the Pre-tests (Tables VI and VII, Figures 5 and 7) *cannot be attributed to any true treatment effect on generalized learning or even on specific improvement*. The inferiority of the SD subjects could be attributed to the slight deficiencies in their equipment. The superiority of the AF group could be attributed to the fact that they took the Pre-tests (and Post-tests) with activated headphones, whereas the others took them with inactivated headphones. The slight superiority of LD over IF is probably best attributed to a random group selection effect which will be discussed later. Indeed, it is not outside the range of possibility that all differences were due to this random effect.

The Pre-tests could justifiably have been employed as aptitude tests relative to the Post-tests as criterion tests to measure the effect of the treatments on immediate improvement of specific utterances. If this had been done, it is obvious from the consistency of the relative treatment standings from Pre-tests to Post-tests (Tables VI and VII, Figures 5 and 7) that no appreciable differences would have been found in the adjusted means of the treatment groups. In short, the experiments *show no superiority or inferiority on the part of any of the treatments in producing immediate improvement through practice on specific utterances*.

The Final Test was administered, like the Aptitude Test, with activated headphones for the first half of the utterances and inactivated headphones for the second half. The AF group showed no superiority on the practiced utterances which were administered with the inactivated condition but did show a statistically non-significant superiority on the practiced utterances administered with the activated condition (Table VIII). *This* *retention of superiority on practiced utterances tested under the activated condition is the only indication of any superiority for the AF treatment as a training device*; and, since it is statistically non-significant, no conclusion can be made concerning even this limited advantage to the AF condition.

On the unpracticed sentences in the Final Untrained, the advantage of the AF treatment was so small as to be negligible. Indeed, the AF group fell below the combined IF and LD averages on the part of the test administered with inactivated headphones (Table VIII). For a laboratory training device to be really superior for teaching pronunciation, its effect must generalize beyond the training situation and beyond the material practiced. The unpracticed part of the Final Test (Final Untrained) provided for only a slight amount of generalization -- to different, but similarly structured, utterances. The only difference found was a statistically non-significant tendency to do slightly better under the condition of activation or inactivation that was present in practice.

The low standing of the LD group on the Final Untrained, when taken into consideration with other data, does indicate a possible inferiority of long delayed playback for generalized learning. This matter will be discussed further below.

In spite of these negative findings, much information has been gained in the course of the experiment which leads to important hypotheses concerning both the measurement of pronunciation performance and the processes of learning to pronounce. Since the experiment was designed to test the effectiveness of the four treatments, the design did not necessarily provide properly controlled tests of the "by-product" findings. It will, nevertheless, be useful to point them out as guides to further research. The "by-product" findings, together with the finding of possible inferiority of the LD condition for "generalized learning" will be stated as hypotheses, followed by discussion referring to the evidence for them.

Hypothesis 1: *In the initial stages of learning to pronounce classroom instruction produces rapid improvement*. *This improvement is large compared with that secured from later laboratory practice*.

Each day the subjects were instructed in the classroom, largely on the general muscular set required for pronouncing French and the general differences between French and American pronunciation. The sentence material for the day was repeated only a few times. Yet the improvement from the Aptitude Test to the Pre-tests was twice that from the Pre-tests to the Post-tests. Since improvement on the Pre-tests was no greater for the latter days of the training period than for the earlier days, it cannot be attributed simply to a generalized improvement as the training period progressed.

The chief objection to taking this result at face value is that, in the course of about a half hour of training (fifteen minutes in the classroom, eighteen minutes in the laboratory) it might be expected that early training would achieve the greater quantitative advance. There is need for experimentation that examines the effects of classroom instruction and laboratory training under directly comparable circumstances.

Hypothesis 2. The Pre-tests, Post-tests and Final Trained measured a capacity for improvement in pronouncing specific utterances which was not measured by the Aptitude Test. This improvement factor was randomly distributed among the treatment groups from experiment to experiment and brought about random variations in their adjusted means from experiment to experiment.

In the discussion of Table D, Appendix IV in the section on results (p.41) it is pointed out that the correlations between the Aptitude Test and the Final Untrained are higher than those between Aptitude and Pre-tests, Post-tests and Final Trained. The obvious difference between the Aptitude and the Final Untrained on the one hand and the Pre-tests, Post-tests and Final Trained on the other is that the subjects had no opportunity to improve their pronunciation on specific utterances on the first two, whereas laboratory practice and/or classroom instruction intervened to improve pronunciation on the last three. A *specific improvement factor* may be postulated which was measured by the tests directly affected by training but not by the Aptitude Test. This factor might actually be the converse of a capacity for quick readiness to do well on the Aptitude Test or it might be a factor of motivation or ability leading to superior performance after practice.

In any case, the improvement factor would be randomly distributed in varying amounts in the treatment groups and would not be accounted for in adjusting the means; hence, treatment groups high in the factor would receive spuriously high means and *vice versa*.

Random variations in this factor would account for the variations from experiment to experiment on the part of treatment groups exhibited in Table VI and Fig. 6. Thus, relative to one another, the IF subjects in the fourth and seventh experiments were "good improvers", whereas in the fifth experiment they were very "poor improvers".

The Aptitude Test was selected as putatively the best predictor of later performance because it constituted an exact work sample of the material that was to be practiced. It would appear, on the basis of results, that its failure to predict the full potential for improvement with practice calls for a different kind of work sample, one in which the utterances selected for scoring occur at a later stage of practice on a given sentence and probably one in which there has been some classroom instruction on how to pronounce. Such a test would be essentially equivalent to the "training criterion tests" of Table D, Appendix IV. Although these tests did not predict the "generalized learning" criterion (that is, the Final Untrained) better than the Aptitude Tests, it should be realized that not much generalized learning appears to have occurred during the six days of Training Sessions. The scores on the Final Untrained do not reach the level of those on the Pre-tests. It might be expected that over a course of several months training, the "improvement factor" would contribute considerably to generalized learning, and an aptitude test containing the "improvement factor" would prove a better predictor; that is, a better measure of aptitude for generalized learning than the test actually employed in these experiments.

Hypothesis 3. The Overall variable alone probably constitutes an adequate measure of pronunciation ability and has the advantage of being more economical than the combined Overall and Phonemic variables.

The experimental results were so similar for the Overall and Phonemic variables and the two variables correlate so closely that there seems to have been little need to employ both variables for any other purpose than mutual corroboration.

Half the time spent in scoring student's utterances was spent on the Phonemic variable. Although this variable was a better predictor of criterion tests than the Overall (Table D, Appendix IV), the Overall after training predicts the Final Untrained better than the Phonemic.

It seems reasonable to assume that after several months of training the Overall variable, tested after adequate instruction and practice, would show a definite superiority to the Phonemic as a predictor of generalized learning. It has the advantage of greater reliability, probably because it offers the raters a larger sampling of a subject's pronunciation skill for each judgment. It also has the advantage of testing intonation as well as accuracy in pronouncing single phonemes.

Hypothesis 4. Pronunciation performance is better under conditions of activated feedback than under conditions of inactivated feedback. However, there is no evidence that this superiority results in a retained superiority in pronunciation in other situations. There is some evidence that subjects who have practiced with activated feedback are superior in performance to subjects practiced with inactivated feedback when both are working with the activated feedback condition.

This hypothesis has already been discussed in the introductory part of this section. Its validity rests upon a decision as to whether the consistent superiority of the AF condition for the Pre-tests, Post-tests and Final Trained on both the Phonemic and Overall variables was due to random selection for the "improvement factor" or to the fact that the AF condition provides superior cues for guiding pronunciation. The consistency of the results from test to test and for both variables precludes the possibility that the differences between treatments could have been produced by random errors in measurement, but it is not beyond the range of probability that the chances of selection could have led to a series of three AF groups relatively high in improvement capacity for each of the three experiments.

The analyses of covariance, shown in Tables I, K, and L, Appendix IV, are designed to put this hypothesis of random group selection to the test. The rejection of that hypothesis rests upon the $P < .05$ level of significance found in four of the analyses. The fact that the data did not meet the test of homogeneity of variance could mean that the F's are either too high or too low, and makes the full acceptance of the reality of better performance with AF somewhat more shaky than the rather low confidence level of $P < .05$ already renders it. However, the evidence strongly favors the acceptance rather than the rejection of the hypothesis of a true advantage to the AF condition in controlling pronunciation performance.

It might be supposed that a more direct measure of the differences between the AF and IF conditions for performance could be obtained by finding the difference on the Aptitude Test, where the first half was administered with activated headphones and the second with inactivated headphones. The difficulty is that this test was not set up for an experimental comparison. The sentences in the first part might have been easier or harder than those in the second part. A priori the time arrangements might be expected to favor the inactivated portion, since it was administered after an 8-minute rest which would eliminate fatigue effects and it thus had the advantage of practice effects which might be considerable with subjects who had never had any previous practice.

In spite of the probable difficulties in interpretation, the data were analyzed. They are not reported in the results section because they were found to be uninterpretable. The activated half of the Aptitude Test yielded a combined score 16 points higher than the inactivated part, a difference significant at $P < .001$. However, the sentences selected for training from the activated part of the Aptitude Test yielded combined scores 18 points higher on the Pre-tests and 20 points higher on the Post-tests than those selected from the inactivated part. Hence, the superiority of the activated section of the Aptitude Test might be due to the fact that the sentences were easier, and no conclusion can be made.

Hypothesis 5. Long delay playback is inferior to a comparably administered non-playback condition in effectuating generalized learning.

The evidence for this hypothesis is indirect and is not statistically significant, but it constitutes the only approach to a positive conclusion that the experiments yield with regard to the effectiveness for learning of the playback conditions. (The complete indeterminateness of the findings with regard to SD will be discussed in connection with Experiment 6.)

The hypothesis is based on the assumption that the consistent superiority of the LD treatment over the IF treatment for the Pre-tests, Post-tests and Final Trained was a result of random group selection on the "improvement factor." In the first place, there was no experimentally varied condition that can account for the superiority of LD on the Pre-tests. Both groups took the Pre-tests with inactivated headphones prior to any experimentally differentiated practice on the utterances. To be sure, the later Pre-tests were taken after experimental differentiation, but the superiority of LD on Pre-tests was not the effect of learning over the six days of practice as evidenced by the fact that the LD group increased its Pre-test score for the last three days over the first three days no more than the IF group did (Table H, Appendix IV).

Furthermore, reference to Fig. 6 shows no consistent superiority of the LD group from experiment to experiment. If only experiments 4 and 7 had been performed, LD would have averaged lower than IF. The extremely low standing of the IF group in Experiment 5 alone accounts for the average superiority of LD over IF for all three experiments. It may be positively concluded that no result of the experiments indicates any superiority for long delay playback over non-playback conditions.

Assuming that the LD superiority on the Pre-tests, Post-tests and Final Trained was a function of chance superiority of the LD groups on the "improvement factor", the fact that the adjusted LD mean falls below the adjusted IF mean on the Final Untrained, offers indicative evidence of inferiority on the part of LD practice to generalize to unpracticed material.

If the Pre-tests and Post-tests had been used to predict achievement on the Final Untrained it is evident from examination of Fig. 7 that the adjusted means of AF and LD would have fallen considerably below the adjusted mean for IF. The effect on the AF mean could be considered spurious because of the probable advantage to the AF condition on the Pre-and Post-tests. But, since LD had no such advantage, the low LD standing could be attributed to a true failure of the LD treatment to produce as much generalized learning as the IF. It is highly improbable, however, that the difference would be statistically significant, hence the assumption of inferiority on the part of the LD treatment for generalized learning must be viewed as definitely hypothetical.

Hypothesis 6. Whatever special difficulty older students experience in learning to pronounce is more a function of inability to retain good pronunciation than inability to achieve it.

This hypothesis is based on the marked and statistically significant difference between the college group and the high school groups on the Final Trained (Tables I, J, and K, Appendix IV, Table VII, and Fig. 7). Examination of Table E, Appendix IV, reveals that the college group received higher scores on the Post-tests than did either of the other groups, but it regressed more on the Phonemic scores in the Final Trained and also regressed on the Overall whereas the other two groups improved. The Combined raw score for the high school groups on the Post-tests was 845, on the Final Trained 846. The college groups scored 890 on the Post-tests and 853 on the Final Trained.

The fact that the significance of the difference between the high school and college groups lies at the $P < .01$ level for the Combined variable leaves little doubt that the college group really did retain less of the improvement it achieved in the Training Sessions.

The college group scored higher on the Carroll-Sapon and the Aptitude Test than either of the high school groups. But on the Final Trained, this advantage disappeared. Thus it might be expected that over the course of a few months of teaching, the high school groups, with better retention of whatever progress they made as a result of instruction and practice, would gradually outstrip the initially more able college group. This effect would, of course, account for the common observation of language teachers that younger students learn to pronounce more readily than their elders.

Note: The term "anchorage" is based on the concept that given qualities of pronunciation are "anchored" to the standard positions on the rating scale in the judgments of the raters. Intermediate qualities are rated on the degree to which they approximate these anchorage points. A given quality for rater B had a consistently higher anchorage point on the scale than for rater C, and rater B's anchorages became higher, relative to C's, as the work of rating progressed.

The Continuation Experiment (Experiment 6)

Results

In Experiment 6, the subjects were trained on both the trained and untrained sentences that the same group had been tested on in Experiment 5. Since each group of sentences had received different experimental manipulation, the results on the Experiment 5 trained sentences were analyzed separately from the results on the Experiment 5 untrained sentences.

Since in the analysis of the Replication Experiments no really important differences between the Phonemic and Overall variables appeared, this report on the Continuation Experiment will be centered on the Combined scores as the probably most reliable index of performance. However, to show the irregularities in the means attained by individual treatment groups from test to test, the raw score means are given for both Phonemic and Overall in Table G, Appendix IV. The Combined means are shown in Table IX. The irregularities appear to be quite uninterpretable and are probably due largely to random variations in scoring or the day-to-day reactions of the subjects. Fig. 8 displays the data on the bottom line of Table IX. The changes in mean Combined scores from test to test are shown separately for the sentences on which the subjects were trained in 5 and those on which they were not trained. The most remarkable result is the failure of the group ever to attain the proficiency in Experiment 6 that it attained on the Post-tests in Experiment 5. The difference between the Post-tests in Experiment 5, and the Post-tests on the same items in Experiment 6 is 25 units, which is significant at the $P < .01$ level ($t = 2.90$, $df = 26$).

All mean differences between treatment groups in Experiment 6 for both Trained-in-5 and Untrained-in-5 sentences were tested for significance for the Phonemic and Overall variables separately and also for the Combined variable. None were found significant. To show similarities or discrepancies in the standings of the treatments from Experiment 5 to Experiment 6, the means of the adjusted Phonemic and Overall scores by treatment are shown for both experiments in Table X. (Adjusted Combined scores are not used because they were not computed for Experiment 5). Mean deviations of the treatment means from the mean of all four are shown to indicate the greater variability between treatments in the Pre- and Post-tests of Experiment 5.

The deviations of the treatment means of Table X from the mean of all four are portrayed in Fig. 9. Major consistencies are the first rank for AF on Pre-tests, Post-tests and Final tests in both experiments and the very low standing of IF on all but the Final Untrained of Experiment 5 and the Introductory and Pre-tests of the Untrained-in-5 sentences in Experiment 6, where SD ranks fourth. Otherwise, SD ranks relatively higher in these two experiments than in Experiments 4 and 7. (Compare Table VI and Fig. 6).

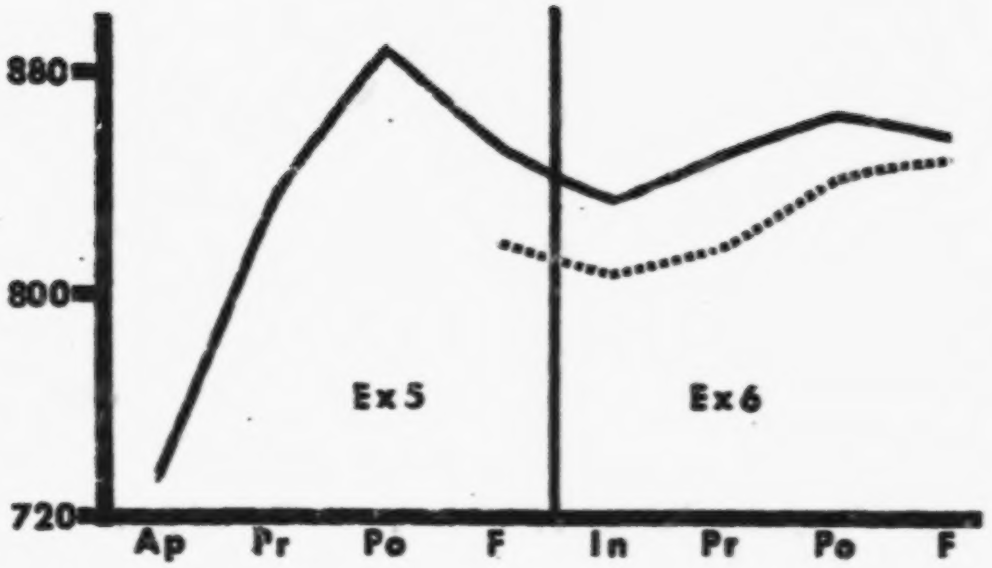

Fig. 8. Means of the college group on the Aptitude Test and all criterion tests in Experiments Five and Six. Solid line: Sentences used for training in Experiment Five. Dotted lines: Sentences not used for training in Experiment Five. (From lower row, Table IX).

TABLE IX

COMBINED MEANS OF PHONEMIC AND OVERALL MEANS OF TREATMENTS FOR ALL TESTS IN EXPERIMENT FIVE, THE PART OF EXPERIMENT SIX TRAINED IN EXPERIMENT FIVE, AND THE PART OF EXPERIMENT SIX NOT TRAINED IN EXPERIMENT FIVE WITH MEANS OF THE TREATMENT MEANS.

Note: In Experiment Six "In" signifies Introductory Test and "F" Final Test.

	Experiment 5					Experiment 6 Trained in 5				Experiment 6 Untrained in 5			
	Ap	Pr	Po	FT	FU	In	Pr	Po	F	In	Pr	Po	F
IF	745	797	852	824	834	800	819	853	838	808	820	830	829
AF	715	852	915	864	819	839	858	882	862	804	824	867	860
LD	732	837	900	856	821	855	858	864	865	825	827	835	849
SD	743	833	843	869	804	839	862	861	865	792	802	840	853
Mm	733	830	888	853	820	833	849	865	858	807	818	843	848

TABLE X

COMBINED MEANS OF ADJUSTED PHONEMIC AND OVERALL MEANS FOR ALL CRITERION TESTS IN EXPERIMENTS FIVE AND SIX, MEANS OF PARTS TRAINED AND NOT TRAINED IN FIVE SHOWN SEPARATELY, WITH MEAN DEVIATIONS OF THE TREATMENT MEANS FROM THE MEAN OF ALL FOUR.

Note: In Experiment Six "In" signifies Introductory Test and "F", Final Test.

	Experiment 5				Experiment 6 Trained in 5				Experiment 6 Untrained in 5			
	Pr	Po	FT	FU	In	Pr	Po	F	In	Pr	Po	F
IF	791	836	818	827	795	814	848	832	798	812	824	823
AF	863	925	874	832	846	867	890	874	825	836	878	870
LD	839	901	857	822	855	858	863	865	827	828	836	850
SD	827	888	866	798	837	859	858	862	783	796	835	848
MD	21	26	18	11	19	18	13	13	18	14	18	12

To determine whether the high standing of the AF group on the Final Test of Experiment 6 was produced by its superiority on the activated part of the test, the means for the AF group on the activated and inactivated parts were compared with the means for the other three groups combined. The AF group scored 872 on the activated section and 850 on the inactivated. The other three groups scored 837 on the activated and 862 on the inactivated. Hence, as in the Replication Experiments, the superiority of the AF group on the Final Test is almost entirely a function of its superiority on the activated part of the test.

Discussion

Experiment 6 was deliberately designed to be purely exploratory. The experimenters wished to see what would happen if training were continued for another six days without classroom instruction and what relationship would appear between the material already practiced in Experiment 5 and the material on which the group had not been trained in Experiment 5. The aim was not so much to obtain definitive results as to secure information on the basis of which hypotheses could be formed.

It should be remembered that the college group employed in this experiment was poor at retaining the skills it had achieved in the Post-tests. If one of the high school groups had performed in this Continuation Experiment, much more improvement

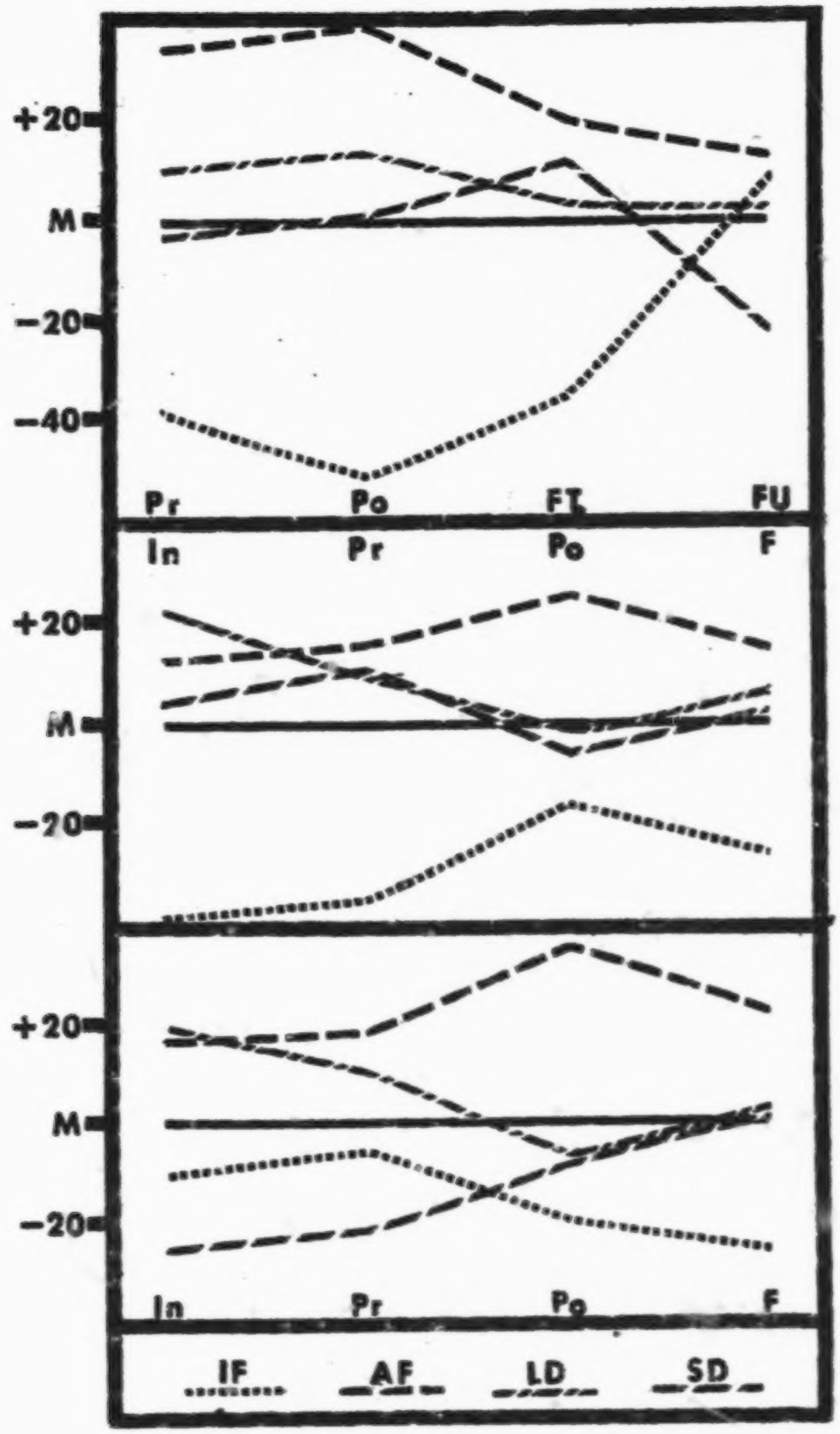

Fig. 9. Deviations of adjusted treatment means from means of all four on criterion tests. Top: Tests in Experiment Five. Middle: Tests in Experiment Six, sentences used in training in Experiment Five. Bottom: Tests in Experiment Six, sentences not used in training in Experiment Five. (From Table X).

on the basis of what had already been learned might have occurred.

Study of Fig. 8 shows that a considerable general improvement in pronunciation occurred during the six days of training in Experiment 5. The scores on the untrained material on the Final Test of Experiment 5 are almost as high as those achieved on the Pre-tests in that experiment. In the course of the week that intervened between Final Test in 5 and the Introductory Test in 6, only a slight decrement occurred. Scores on the Pre-tests, Post-tests and Final show that training on the formerly untrained items brought about gradual improvement, but on these items the subjects never reach even the level that they attained on the Final Trained in 5.

On the utterances which were practiced in Experiment 5, the subjects maintained a somewhat higher average. Their scores on these utterances dropped below the level of the Experiment 5 Pre-tests in the Introductory Test of 6 and on the Final Test reached only the level of the Final Test in Experiment 5. On the Post-tests, they failed markedly to reach the level of the Post-tests in Experiment 5.

In short, a second six days of training not only failed to produce improvement over the level attained in Experiment 5, it actually failed to achieve the level in the Post-tests that was achieved in Experiment 5. The possibility of a downward shift in the anchorages of the raters cannot be excluded, but it seems unlikely, since the chief change in anchorages in the course of the Experiments would appear to be an upward shift on the part of Rater B. (See Table B, Appendix IV). Actually, in Experiment 6, the Phonemic scores show even less improvement than the combined scores do because there is a greater difference between Overall and Phonemic scores in 6 than in 5. This would fit the assumption that Rater B's anchorage point rose steadily throughout the four Main Experiments and that the scores in Experiment 6 are spuriously higher than they should be relative to Experiment 5.

Another possibility is that the practice of four sentences during the laboratory Practice Sessions made it difficult to achieve the high level of excellence that could be attained when concentrating on only two utterances.

A third possibility is that the absence of the Classroom Session resulted in failure to re-establish the set toward good pronunciation and hence made it less possible to improve. The subjects may have forgotten the postures for correct French pronunciation that they had been reminded of daily in Experiment 5, and have tended to revert throughout the Practice Session to American speech postures. In view of the apparent marked effect of the Classroom Session of producing great improvement in the Pre-tests over the Aptitude Test it seems reasonable that this was at least one of the factors involved. If the absence of classroom instruction was a major cause of lack of improvement in Experiment 6, it may be concluded that classroom instruction can not only bring about improvement, but that it can set the stage for effective improvement in laboratory practice. Perhaps its most important effect could have been to remind the subjects of the vocal posture required for effective French pronunciation.

In comparing the adjusted means, it should be remembered that none of the differences shown are statistically significant. Nevertheless, if these differences fit the hypotheses formed on the basis of the Replication Experiments, they provide additional support for those hypotheses. Such a comparison fits the hypothesis that the AF group received advantage from taking the Pre-tests and Post-tests with activated headphones and that this advantage carried over to the activated section of the Final Test.

First let us consider the AF performance on the part trained in Experiment 5. (See Fig. 9). On the Final Trained, AF lost much of the superiority it had displayed on the Pre-tests and Post-tests because it retained superiority only on the activated parts. On the Introductory Test of Experiment 6 it had lost still more. On the Pre-tests it again takes first place, but is not far ahead of LD and SD. This may be because the absence of a Classroom Session lessened the degree of advantage to the AF group of the superior cues which the activated condition presumably provides. On the Post-tests, however, AF is as much superior to the next best treatment as it was in Experiment 6. Again on the Final Trained, the AF superiority declines, and it has been shown that this decline is due to its inferiority on the unactivated part of the Final Test. Practically the same pattern is repeated on the material not practiced in Experiment 6, except that here the relative superiority of AF on the Post-tests is even more marked. The outcome of Experiment 6 thus supports the fourth hypothesis derived from the Replication Experiments.

The relationships between the SD and IF treatments conform to the hypotheses that there were group differences in improvability which were not measured by the Aptitude Test and that the SD treatment group was handicapped by inferior sound production in their machines.

Throughout the running of the experiments, the experimenters did not realize that what seemed to them to be a relatively slight inferiority in the SD machines might be a genuine handicap. They therefore allowed the subjects to sit where they happened to go for the Aptitude Test and assigned them to the special treatment positions on the next day. Assuming that the SD equipment was a handicap, chance differences between groups in the number of members seated at the SD booths may have resulted in relatively low Aptitude scores in some treatment groups and relatively higher criterion scores, thus contributing to making certain groups "better improvers" than others. The procedure would, however, tend systematically to lower criterion scores relative to aptitude scores for the SD groups, and could account for their generally low adjusted means in the Replication Experiments.

In Experiment 5, the SD group performed relatively better than it did in the other experiments, whereas the IF performance was relatively much poorer except on the Final Untrained. (See Table VI and Fig. 9). In terms of Hypothesis 2 in the discussion of the Replication Experiments, the SD subjects in Experiment 5 were "good improvers" and the IF subjects "poor improvers". SD did well, IF poorly where they had previous classroom instruction for the Pre-tests or previous practice on the material for the Post-tests and Final Trained. On the Final Untrained, without previous instruction or practice on the material, the handicap under which the SD group worked reduced it to last place; whereas the IF group made a score essentially equivalent to its standing on the Aptitude Test.

In Experiment 6, SD maintained its relatively high position near the mean of the group on the sentences that had been trained. On the untrained sentences, however, it ranked below IF on the Introductory Test and Pre-tests, both of which were presented prior to any classroom instruction or training on the material. In short the "good improvers" in the SD group do poorly without instruction and practice because of their handicap. The "poor improvers" in the IF group do poorly where opportunity for improvement has been given.

To sum up, the results of Experiment 6, although not in themselves statistically significant, conform to the following assumptions derived from analysis of the Replication Experiments:

1. Superiority of AF on the Pre-tests and Post-tests indicates that the AF condition provides superior cues for pronunciation performance.

2. Retained superiority on AF generalizes no further than to superiority under the activated condition on material already practiced under that condition.

3. Certain treatment groups in each experiment displayed relatively high or low special ability to improve that was not measured by the Aptitude Test.

4. The SD treatment group was handicapped on the criterion tests, presumably because of inferior sound production in their machines.

The Trial Experiment

Results

The procedure in the Trial Experiment was closely similar to that in the Replication Experiments. The results, therefore, have been analyzed to further test the hypotheses developed on the basis of those experiments and to discover what differences might have been produced by minor differences in procedure.

Table XI is comparable with Table V and Table E, Appendix IV, and of the Replication Experiments. There is less difference between Phonemic and Overall scores in the Experiment 3 data, but, as has already been shown, (Table B, Appendix IV) this may well be due to changes in the anchorage points of the raters. In any case, both variables show the same pattern of gain and loss.

The subjects in Experiment 4 and 7 were in the same age range as those in Experiment 3. These subjects in the Replication Experiments gained somewhat more, especially in the Pre-tests and Post-tests.

An analysis of covariance for each criterion test was performed for both the Phonemic and Overall variables without any finding of significant differences. The adjusted means for the three treatments and the deviations from the mean of all three are shown in Table XII, comparable with Table VI and Table G, Appendix IV. None of the treatments in Experiment 3 show any consistent superiority. This is emphasized by the very slight differences between the means of all tests and both treatments in the right hand column of Table XII.

Discussion

The lower gains in the Trial Experiment may be attributed to the fact that the students worked in the laboratory for three days before taking the Aptitude Test, hence they had already made some progress. The fact that their standing on the Pre-tests and Post-tests was lower relative to the Final Test may be attributed to the fact that the former were not emphasized as being tests.

The failure of the AF treatment to exceed the others on the Pre-tests and Post-tests may be due to the fact that the AF feedback was of inferior quality in Experiment 3, as noted in the section on equipment. Indeed, this inferiority was more noticeable to the experimenters than the inferiority of the SD equipment in the Replication Experiments.

TABLE XI

MEANS OF TREATMENTS IN THE TRIAL EXPERIMENT FOR PHONEMIC AND OVERALL VARIABLES WITH DIFFERENCES BETWEEN APTITUDE AND CRITERION SCORES AND WITH MEANS OF BOTH THE MEANS AND THE DIFFERENCES.

		N	Ap	Pr	Po	FT	FU	DPr	DPo	DFT	DFU
	IF	8	632	738	767	794	721	106	135	162	89
	AF	8	648	708	756	786	726	60	108	138	73
Ph	LD	8	666	732	767	782	721	66	101	116	55
	Mm		649	726	763	787	723	77	114	138	74
	IF	8	697	732	777	861	730	35	80	164	33
	AF	8	723	785	826	882	765	62	103	159	42
OA	LD	8	724	786	851	867	757	62	127	143	33
	Mm		715	768	818	870	751	53	103	155	36
CoMm			682	747	791	829	737	65	109	147	55

TABLE XII

ADJUSTED MEANS OF TREATMENTS IN THE TRIAL EXPERIMENT FOR PHONEMIC AND OVERALL VARIABLES AND COMBINED MEANS FOR EACH CRITERION TEST WITH MEANS OF MEANS FOR EACH TREATMENT, MEANS OF MEANS FOR EACH TEST, AND DEVIATIONS OF TREATMENT MEANS FROM THE MEANS OF THE TEST.

		Pr		Po		FT		FU			
		M	D	M	D	M	D	M	D	Mm	D
	IF	751	+25	781	+18	806	+19	731	+8	767	+17
Ph	AF	709	-17	756	- 7	786	- 1	726	+3	744	- 6
	LD	718	- 8	753	-10	770	-17	711	-2	740	-10
	Mm	726		763		787		723		750	
	IF	744	-24	787	-31	873	- 7	744	-7	787	-15
OA	AF	780	+12	821	+ 3	877	+ 7	758	+7	809	+ 7
	LD	781	+13	846	+28	861	- 1	750	-1	810	+ 8
	Mm	768		818		870		751		802	
	IF	748	0	784	- 7	840	+11	738	+1	777	+ 1
Co	AF	745	- 3	789	- 1	832	+ 3	742	+5	777	+ 1
	LD	750	+ 2	800	+ 9	816	-13	731	-6	774	- 2
	Mm	748		791		829		737		776	

In short, the results of the Trial Experiment neither support nor contradict the findings of the Main Experiments and point to little but the fact that relatively minor differences in procedure may lead to definite differences in results. Assuming superiority in performance for the AF condition, they conform to the hypothesis that minor deficiencies in sound quality may handicap accuracy of mimicry.

General Discussion and Suggestions for Further Research

Activated vs. Inactivated Feedback

The major positive finding of this series of experiments is the probable superiority of the activated headphone condition to the inactivated condition for pronunciation performance. Although measures of the statistical significance of this finding are marginal, the result may easily be confirmed or disconfirmed by a relatively simple experiment specifically designed to compare performance only without concern for learning. An important requirement for such an experiment should be that the sound quality of both the model's utterance and the activated feedback should be high and equally high.

The failure of the AF groups to display any appreciable retained superiority on the Final Test, except for the material which they had already practiced and which was presented in the test with activated headphones casts considerable doubt on whether or not students would actually learn to pronounce better with activated headphones. Even if the activated headphone condition does present better cues for controlling performance, there is no a priori reason for believing that a long term of practice with this condition would result in generalized learning of better performance in the normal conversational situation. The absence of the cues under which practice took place might actually serve as a handicap, since learned skills are notoriously dependent on the cues under which they have been learned. The fact that the AF groups did less well than the other groups on the part of the Final Untrained administered with inactivated headphones in the Replication Experiments and similarly on the Final Test in the Continuation Experiment is indicative of this possibility.

As pointed out in the introductory analysis of the problem (p.3), activated feedback, with a relatively low gain setting, produces an impression similar to ordinary speech conditions. Practice with such a setting might be optimal for generalization from laboratory practice to the conversational situation.

The above considerations are, of course, speculative. It might be that the condition which produces the best performance would, over the course of time, produce the highest degree of generalized learning. Probably the best means of testing would be to equate three groups in several language classes, let each group, over the course of a year, practice solely with one of three feedback conditions: (1) inactivated headphones, (2) activated headphones with gain set to reproduce normal speech conditions, (3) activated headphones with gain set about as high as is comfortable. Then compare the groups by testing in a conversational situation.

Delayed Playback

Some language teachers express the opinion that long delay playback bores students and is a waste of time. The questionnaire, however, failed to show greater boredom in the LD group, but the technique used in these experiments of having students repeat utterances while listening to the playback may have alleviated boredom as well as straying of attention. At any rate, the LD groups showed as much improvement on specific sentences in the Post-tests and Final Trained as did the comparable IF groups, although they practiced mimicry only half as much. Possibly because of this truncated practice, the LD groups showed indications of inferiority in generalized learning on the Final Untrained. There was certainly no indication of superior learning resulting from the use of the highly expensive playback equipment.

There can be little question that listening to the playback of one's own voice is interesting, motivating and possibly instructive; but _a priori_ it would appear that if the student listens back to everything he does, the procedure should become quite boring and furthermore divert time from valuable practice. An ideal arrangement might be to have a few recording machines in a laboratory with which students could test themselves by listening back as often as they wished. Such an arrangement could be compared experimentally over the course of a year with arrangements allowing for no playback.

Because of the possible handicap of inferior sound production, the experiments provide no certain evidence in regard to the effectiveness of the short delay playback condition. Further experimental work with this condition should be done, since there are _a priori_ reasons for expecting it to be effective as well as reasons for doubting its effectiveness as outlined on page 7 in the Analysis of the Problem.[1]

---

[1] Professor Rand Morton of the University of Michigan has recently been experimenting with a short delay device under the control of the student which echoes the student's utterance in the same way that it was echoed in our short delay equipment. In a personal communication he states that this device results in more rapid achievement of a criterion of student satisfaction with pronunciation on frames of sixty utterances each and that the degree of pronunciation achievement with which students are satisfied is the same with or without the short delay device.

The students in these comparisons have all had 30 hours of training in discrimination of the target language sounds. This type of training might well improve ability to take advantage of activation as well as short delay playback, and in future studies of the effectiveness of laboratory equipment, the relative value of various forms of equipment with pre-training in discrimination should be investigated.

Sound Quality

The compulsions of scheduling which led to the use of the short delay equipment before it could be thoroughly tested and brought to full equality in sound production with the rest of the equipment were certainly unfortunate from the standpoint of experimental rigor. But the hint offered as to the possible importance of good sound production for securing the best possible results in teaching pronunciation may have valuable repercussions.

As mentioned in the section on equipment, the sound qualities of the equipment in the laboratory were judged by a competent outside observer to be of the highest order. The defects in the short delay equipment were relatively slight, and the short delay sound system was probably superior to the sound systems in many laboratories. Yet there are definite indications, especially in the fluctuations in the standings of the SD group of "good improvers" and the IF group of "poor improvers" in Experiments 5 and 6, that these relatively slight defects were definitely handicapping. The failure of the AF group to demonstrate superior performance in the Trial Experiment where their activated feedback was inferior in quality to the sound of the model's voice is another indication of this possibility.

In the opinion of the experimenters, experimental comparisons of performance using the best producible equipment with equipment of various degrees of inferiority are definitely called for. In their opinion, also, it might be found more profitable for language laboratories to expend funds to bring their equipment to the highest possible level of acoustical excellence than to spend them on devices for activated feedback or playback. Again, however, it should be pointed out that high performance is not necessarily a guarantee of superior generalized learning and that further research may bring evidence for the value of activation and playback that this investigation failed to uncover.

Classroom Instruction

The results indicating the importance of classroom instruction conform to what is generally known about the acquisition of motor skills. Practice, involving self-correction and overlearning, is necessary; but instruction in the right way to effect a skilled performance and correction based on the perceptions of trained observers are important, especially if the highest degree of skill is to be reached. The classroom instructor in these experiments was engaged in the same function as the athletic coach when he describes to his proteges the correct "form" to employ. His function did not, of course, include the function performed in many classrooms of correcting errors.

If improved techniques of teaching pronunciation are to be developed, the best classroom techniques are deserving of study. A major emphasis of the highly experienced instructor in this experiment was stress on the correct "posture" for speaking French. The results of the Continuation Experiment suggests that failure to administer this instruction each day before entering the laboratory resulted in lower improvement for the day.

The function of the classroom instruction can be related theoretically to B. F. Skinner's observations on animal learning. An animal placed in a Skinner box is reinforced whenever it emits a certain response. Soon the reinforced response is regularly emitted; that is, it has become "conditioned". However, if a response not in the animal's repertory is desired, some method of getting the animal to emit the response must be employed before reinforcement can have any effect. The Skinnerian method is to "shape" the response by first reinforcing any part of it or any approximation to it which occurs. When an approximation has been conditioned, only the instances of the conditioned response which more closely approximate the desired response are reinforced. By this method of successive approximation, the desired response is finally shaped and conditioned.

Self-correction of pronunciation in the laboratory is the analogue of reinforcement of emitted responses. The student is expected to increase the frequency of responses which most nearly approximate the model. But students who have never spoken another language do not have the sounds of the language in their repertory. If their responses are not "shaped" in some way, they may never emit responses even closely approximating those of the target language. Some "shaping" is doubtless achieved in the process of self-correction, but among human beings,one of the best ways of shaping responses is to tell the learner how to produce the desired response. This can be vastly more efficient than the slow process of shaping animal responses to which whatever shaping is achieved in the laboratory is analogous.

When a response has been fully conditioned in a Skinner box it can be "extinguished" that is, reduced to a minimum frequency of occurrence, by withholding reinforcement when it occurs. A day later, however, it will appear and require a new extinction. The sounds of a native language have been strongly conditioned. In the course of a session of practice, they become more-or-less extinguished and the sounds of the target language conditioned through the process of self-correction. A day later, however, there is likely to occur a strong spontaneous recovery of the native language sounds. Hence, it would appear that a "reshaping" of the target language sounds immediately preceding laboratory practice would be a daily necessity for a considerable period of time in the course of learning to pronounce, in order to provide for a large number of responses within the range of the target language for reinforcement in the course of laboratory practice.

The failure of the subjects in Experiment 6 to achieve the levels in the Post-tests that they did in Experiment 5 conforms to the above theoretical considerations. An experiment to test the relative advantages of, say thirty-five minutes spent in the laboratory without pre-instruction against twenty minutes in the laboratory with fifteen minutes pre-instruction could test the practical value of the hypothesis that greater progress will be made in conditioning target language sounds and extinguishing native language sounds if some "shaping" is achieved before each laboratory practice.

Enough has already been said about the problem of generalization or transfer from the laboratory situation to the conversational situation to point to the desirability of classroom practice together with laboratory practice in order to insure generalization of the skills achieved in the laboratory to a more conversation-like situation.

None of the above discussion is intended to suggest that classroom instruction alone can provide as efficient learning as classroom work -- and, if possible, individual instruction -- combined with laboratory practice. The laboratory obviously offers the opportunity for more intense and concentrated practice than the classroom, making for rapid extinction of native speech sounds and conditioning of target language sounds, provided the target sounds have been "shaped" so that they occur with considerable frequency.

Standardizing the Testing of Pronunciation

The need for a well-standardized work sample test of pronunciation aptitude is fairly obvious, both for purposes of equating experimental groups and as a means of sectioning classes. Certain requirementsof such a work sample test are applicable to standardized tests of achievement as well as of aptitude. The results of the experiment indicate that the problem of getting reliable ratings is not a difficult one, although it might be necessary to test and/or train individual raters to make certain that they were able to rate as reliably as the raters in this experiment. The lower reliabilities in Experiments 5 and 7 than in Experiment 4 suggest that rater fatigue may be an important factor to watch in any large scale testing.

One definition of validity for a test of pronunciation should certainly be agreement among recognized authorities as to what constitutes good pronunciation. The very high level of agreement between raters found in these experiments -- as far as relative standings are concerned -- suggests that this should not be a difficult problem. But this may be because the raters influenced each other. For a standardized test of pronunciation, it should be demonstrable that recognized authorities, scoring independently, produce highly correlated results.

The wide-ranging anchorages of the raters in these experiments suggest a definite problem for achieving a standardized test of pronunciation. Unless the anchorage of different raters could be somehow standardized, the scoring of a standardized test would give varying averages from rater to rater, no matter how well raters agree as to the relative standings of individuals. Standardization would therefore require a standard sample of scorings with which raters could practice, comparing their own ratings with those of the sample until they found they were rating in terms of the anchorage points of the sample; and they would need to return to the sample occasionally to make sure their anchorages were not "drifting". In short, standardized pronunciation tests would require trained scorers, just as many standardized psychological tests do.

The above considerations apply to all kinds of standardized pronunciation tests. With respect to the requirements of a standardized aptitude test, the chief finding from the experiments is the desirability of measuring pronunciation after some opportunity for improvement. The high correlations of the Pretest scores with the other "training criterion tests" in Table D, Appendix IV suggest that the "shaping" derived from pre-instruction is the major factor producing the condition of improvement over the more naive approach to sheer mimicry which characterized our subjects in taking the aptitude test. Actually, some experimental work would be required to determine the best methods of pre-instruction and the amount of practice on sentences requisite to a stable measure of "improved" pronunciation. Such a standardization of measures of aptitude would obviously require some training of instructors as well as raters.

Finally, our findings strongly suggest the superior efficiency and economy of confining the scoring to an overall rating of whole utterances. Such ratings should be more valid than the scoring of single phonemes, since they would involve judgments of intonation as well as phonemic accuracy.

Age Level and Pronunciation Aptitude

The comparison between the college and the high school groups provides an interesting bit of information relative to the common observation that the capacity to achieve good pronunciation decreases with age. Under practice conditions, the college group actually performed better than the high school groups, but on the Final Trained, the measure of retained improvement, its scores were markedly low *relative to its initial aptitude*. (The fact that its average raw score was slightly higher than the average of the two high school groups is uninterpretable because of uncertainty as to the anchorages of the raters.)

The results of the questionnaire indicated that the college students worked conscientiously in the practice sessions, but that the work was not as meaningful or enjoyable to them as it was for the younger subjects. One might speculate that, although they learned to emit good responses, they were not as strongly reinforced for doing so. Furthermore, the longer years of practice of their own language, or an age-correlated physiological change in flexibility of the learning process, might have resulted in a more massive spontaneous recovery of native language responses.

Perhaps the chief impact of the finding, assuming that it is fully confirmed by later research, is the suggestion that adults are fundamentally as capable of achieving good pronunciation in a target language as are children. Common observation suggests *that they almost never do*. The explanation might be as follows: First, adults need to work longer to achieve good pronunciation, and they are less spontaneously inclined to engage in such work than children. Second, adults learn to talk a language fluently before they achieve a near approximation to native pronunciation, both because they are slower than children in learning to pronounce and because they may learn fluent speech more rapidly. Third, once the individual begins to speak fluently, the inferior pronunciation is over-practiced, and it becomes extremely difficult to improve it.

Wherever it is desirable to develop near-native pronunciation in adults beginning the study of a foreign language, the feat might be accomplished by requiring them to practice pronunciation, always with adequate instruction, for a considerable time before attempting to engage in active speech or conversation. During that time they could be learning the vocabulary and structure of the language through reading. But active speech might well be delayed until the highest possible pronunciation skill has been achieved and well over-learned.

The foregoing discussion should suggest, at any rate, that the common observation that older students do not learn to pronounce as well as children offers a challenge to investigation of the actual factors producing the effect and raises the question as to whether methods of teaching can be developed and tested that may overcome the special difficulties that older students face, whatever they may be. The finding that one of these difficulties may be a handicap in retaining the new skills once they are achieved appears to the experimenters to be an important step in this direction.

Types of Research on Pronunciation

At several points in this report, emphasis has been placed on the fact that no conclusion can be made with regard to the efficiency of a method of teaching pronunciation without testing it out in actual language courses and with performance in actual conversational situations used as the criterion of learning. The latter requirement is important, since there is no certainty that learning with a device or procedure which works well in the laboratory will generalize effectively to the situations in which the language will actually be used. Experiments continuing throughout a course are desirable, since the differential effects of two laboratory devices may only slowly effect differentiation in generalized learning. But they are also desirable because minor variations in experimental procedure may easily change the apparent outcomes of two experiments, and in the language course, the situation resembles most closely the practical situation in which the compared methods and devices are to be used.

A case in point is that motivation in a specialized experimental situation is usually higher than in day-to-day classroom work. Differences in motivation can have major effects on learning. High motivation in all experimental groups can mask out differences between treatment methods which might display true differences under more relaxed conditions. This is particularly true, of course, if one treatment is intrinsically more enjoyable or interesting than another. It might well be that the activated and playback conditions would show superiority in actual course work simply because, as indicated in the questionnaire, students are interested in hearing their own voices.

Another point at which real treatment differences might be masked out by an experiment directed solely to testing pronunciation, is that such an experiment centers attention on pronunciation. Where everyone is trying his best to pronounce, the results may well differ from those that might arise in a situation where attention is more divided. Here again, the activated and playback conditions might show a superiority in a course situation which would not appear in an experimental situation because they might tend to call attention to pronunciation more than the inactivated condition.

The specially designed experiment can never really answer the question of what is actually going to happen in the practical situation. Its function is to tease out the variables which may be important in practice. The present series of experiments was deliberately designed to measure a number of variables, since it seemed to the experimenters that this would be the most economical procedure in the light of the fact that the investigation was opening up a new field. The results raise a number of questions that may be answered in terms of more definitely focussed experimental work. The answers to these questions may be useful guides to the planning of investigations in the actual language teaching situation.

In addition to the suggestions already made along these lines, there is much opportunity for experimental work in the area of programming. What kinds of programs arouse the greatest spontaneous interest and motivation? Is progress more rapid and/or retention better with such programs? Does knowledge of meaning and sight of the written word actually handicap mimicry? If it does so in some circumstances, does it do so in all?

In addition to testing the effectiveness of pre-training in phoneme discrimination, the question may be raised as to whether pre-training in discrimination of intonations might be effective. Furthermore, the question may be raised as to whether a fairly long period of listening to a target language prior to the beginning of mimicry or active speech might not "set the stage" for more rapid progress as well as progress leading to a finally higher level of proficiency.

These and many other questions may be put to the test of specially arranged experiments and finally to the "pay-off" tests of effectiveness in the actual teaching of courses.

VII. SUMMARY

A series of experiments was designed to test the efficiency of the use of four types of language laboratory equipment for learning to pronounce French. The four types were (1) inactivated headphones (2) activated headphones (3) playback after recording a practice session, (4) short delay playback immediately after the recording of a single utterance. (Hereafter referred to as IF, or "inactivated feedback," AF, or "activated feedback," LD, or "long delay playback" and SD, or "short delay playback".)

After preliminary experimentation, three Replication Experiments were performed, each with 7 subjects in each of the treatment groups. The subjects in the first of these were senior high school students; in the second, college students; and in the third, junior high school students. (In the college student experiment one subject dropped out of the SD group.) With the college group, a Continuation Experiment was performed to observe progress over a longer time than that allocated to each of the Replication Experiments.

None of the subjects had ever studied French. Throughout both testing and practice sessions they mimicked a model tape recording. They were not given knowledge of the meaning of utterances until all experimental work was completed.

In each of the Replication Experiments exactly the same procedures were employed. Each experiment began with an Aptitude Test of twenty-four six-syllable sentences. The sentences were built up from one-syllable utterances by gradual addition of syllables to full six-syllable utterances. The first half of the Aptitude Test was administered with activated headphones and the second half with inactivated headphones.

Two of the odd-numbered sentences from the Aptitude Test were used as training sentences each day for six days. On the last day the Aptitude Test was administered again as a final criterion test. The twelve sentences used in training were scored separately from the twelve not used to comprise the Final Trained and the Final Untrained criteria respectively.

Each Training Session began with a fifteen minute period of classroom instruction which was the same for all treatment groups. The subjects were instructed in correct vocal postures for French pronunciation and the training sentences for the day were briefly practiced under instruction. The Classroom Session was followed by eighteen minutes of laboratory practice in which each of the treatment groups worked with its specific treatment condition. Prior to and after each Practice Session, a Pre-test and Post-test was administered which tested pronunciation on the two sentences for the day exactly as it was tested in the Aptitude Test and Final Test. In the Pre-tests and Post-tests, as well as the Practice Session, the AF group worked with activated headphones and the other groups with inactivated headphones.

Analysis of results showed that the Pre-tests constituted a criterion of improvement over Aptitude produced by classroom instruction. The Post-tests measured further improvement on the specifically trained sentences, the Final Trained measured retention of improvement on specific sentences, and the Final Un-Trained alone measured genuine generalized learning to pronounce.

Two variables were measured on each test: (1) Phonemic, measuring accuracy in pronouncing French phonemes particularly difficult for American speakers; (2) Overall, measuring overall correctness in pronouncing three-syllable and six-syllable utterances. The statistical analysis revealed no important differences in results between these two variables, and much of the report is made in terms of a Combined variable produced by combining the scores or means of the two variables.

The following modifications on the above procedures were introduced in the Continuation Experiment: (1) Four sentences were practiced every day, so that all the sentences in the Aptitude-Criterion test were practiced. (2) The Practice Sessions were not preceded by Classroom Sessions.

A questionnaire administered immediately after each experiment revealed no differences between treatment groups with respect to interest and morale. The college group was significantly lower than the two high school groups in interest and morale. All groups claimed to have worked conscientiously throughout and to have tried to do their best most of the time. It was noticed that the junior high school group appeared to enjoy mimicry for its own sake, whereas continuous mimicry was boresome and monotonous to the college group.

The relative achievement of the groups was compared and its significance tested on each of the four criterion tests by means of analysis of covariance. A series of twenty-four analyses for each of the two variables, each test, and each experiment found only one set of significant differences, and this finding was judged to be a random variation. There was considerable variation in treatment standings from experiment to experiment. This was judged to be due to random variations in group selection produced by the fact that the Aptitude Test did not measure a differential factor for improvement over Aptitude standing. In spite of variations in standing from experiment to experiment, the AF treatment averaged high for all experiments and the SD low. SD was dropped from further analysis on the ground that its low standing might be an experimental error occasioned by a slight inferiority in sound quality on the part of the SD equipment. Indirect evidence that the SD group suffered a handicap in taking tests was derived from the data of the Continuation Experiment.

Two-way analyses of covariance were performed for each criterion test comparing standings for the IF, AF, and LD groups and the three experimental groups. On twelve analyses, namely for the Pre-tests, Post-tests, and Final Trained for the Phonemic, Overall, and Combined variables, AF was markedly superior to LD and IF and LD was moderately superior to IF. On the Final Trained AF was superior only on the part of the test administered with activated headphones. It was judged that these consistent differences could be due to random group variations in the differential factor for improvement or to some systematic experimental variable. On the ground of three marginal findings of statistical significance and a fourth significant difference found between AF and IF alone, it was judged that the most probable reason for AF superiority was in part, at least, that AF provided better cues, making for better performance in taking tests. On the basis of convergent considerations, it was judged that the consistent superiority of LD over IF was probably due to random group selection.

On the Final Untrained for all three variables AF stood first, IF second, and LD third, but the variance was small and statistically non-significant. AF was superior to the other two on the part administered with activation but inferior on the part not so administered. It was judged that, in spite of better performance during training, the AF condition did not display superiority under the conditions of this experiment as far as actual learning to pronounce is concerned. It was judged that the decrement in LD performance from the training tests to the Final Untrained might point to a somewhat lower efficiency of the LD condition for generalized learning to pronounce.

It was concluded that the experiment had failed to demonstrate any differences between treatments in efficiency for learning to pronounce except for possible lower efficiency on the part of the long delay condition.

The analyses of covariance revealed a marked and statistically significant deficiency on the part of the college group on the Final Trained. It was judged that difficulties which older students encounter in learning to pronounce may be due more to an inability to retain the results of improvement than inability to achieve good pronunciation in the course of a practice session.

Since the improvement from Aptitude to Pre-tests was about twice as great as the improvement from Pre-tests to Post-tests, and since the college students in the Continuation Experiment fell considerably below the levels on the Post-tests than they had achieved in the preceding Continuation Experiment, the pre-instruction in the Classroom Sessions was judged to be useful, not only for producing improvement, but for setting the stage for later improvement in laboratory practice.

Two results of the experiments suggested that relatively minor deficiencies in the sound quality of laboratory equipment may result in definite lowering of performance. The first was the relatively low performance of the SD group. The second was the failure of the AF group to perform in a superior fashion in a Trial Experiment where the sound quality of the activated feedback was inferior to that obtained in the Replication and Continuation Experiments.

The General Discussion in the foregoing section contains considerable theoretical analysis of the experimental results as well as suggestion as to their relevance to teaching situations. There are also several suggestions as to further research based on the results of the experiments.

APPENDIX I

TESTS AND TRAINING PROGRAMS

Contents

Page

Aptitude-Criterion Test, Main Experiments ------ 81

Training Programs, Replication Experiments ----- 82

Training Programs, Continuation Experiment ----- 85

Aptitude-Criterion Test, Trial Experiment ------ 87

Training Programs, Trial Experiment ------------ 89

APTITUDE-CRITERION TEST, MAIN EXPERIMENTS
(Experiments 4, 5, 6, 7)

The following two sentences show how each sentence in the test was built up. The number (2) after an utterance indicates it was presented twice in succession. Underlined phonemes were scored; always in the second presentation of the utterance. The second presentation of the three syllable utterance and also the second presentation of the six syllable utterance were scored for approximation of the whole utterance to the French phonological pattern.

Un (2)
bon (2)
na (2)
Un bon
Un bon a (2)
Un bon ami
Un bon ami vient
Un bon ami vient tôt (2)

Sur (2)
les (2)
monts (2)
Sur les
Sur les monts (2)
Sur les monts chiens
Sur les monts chiens et
Sur les monts chiens et daims (2)

The following lists the sentences in the test. Target phonemes are underlined.

First Half (Presented with Activated Headphones)

Warm-up sentence, not scored:

Je la vois chaque mois d'août.

Scored sentences, in order of presentation.

(1) Un bon ami vient tôt.
(2) Sur les monts chiens et daims.
(3) Ils sont un peu fanés.
(4) Ôte ce bon preux de l'eau.
(5) Donne-moi sept tapis mauves!
(6) Une jument heurte le sol.
(7) Conduis du bon côté.
(8) Tu as les cheveux plats.
(9) Claude, je veux un stylo.
(10) Sa mule peureuse culbute.
(11) Deux à onze livres cinquante.
(12) C'est la plainte d'aucune sainte.

(Eight minutes rest between First Half and Second Half)

Second Half (Presented with Inactivated Headphones)

Warm-up sentence, not scored:

Le coq jaune chante bien mal.

Scored sentences:

(13) A bas, consul affreux!
(14) Le gamin fin l'atteint.
(15) L'ozone a sauvé Paul.
(16) Julie veut faire la queue.
(17) On a vu neuf bûches jaunes.
(18) Chantons cette belle chanson.
(19) Au printemps il pleut trop.
(20) Il a su le faire seul.
(21) Ton oncle plonge jusqu'au fond.
(22) Jouons bien au ton d'un luth!
(23) C'est combien ce chapeau?
(24) Jules monte et tombe cinq fois.

TRAINING PROGRAMS, REPLICATION EXPERIMENTS
(Experiments 4, 5, 7)

The following shows the program for the first day. The same pattern was followed in all succeeding days. The build-up pattern for the Pre-test and Post-test was the same as in the Aptitude-Criterion Test (q.v.) and both were scored in the same way. The target phonemes are underlined.

After the Pre-test was administered, the Training Series was presented twice, with an eight-minute rest between the first and second presentations. Then the Post-test warm-up and Post-test were given.

The number (2) after an utterance indicates it was presented twice in succession. Target phonemes are underlined.

(Read both columns to horizontal line.)

Pre-test and Post-test

Claude (2)
je (2)
veux (2)
Claude je
Claude je veux (2)
Claude je veux un
Claude je veux un sty
Claude je veux un stylo. (2)

A (2)
bas (2)
con (2)
A bas
A bas, con (2)
A bas, consul
A bas, consul a
A bas, consul affreux! (2)

Training Series, Warm-up Words

dis (2)
tir (2)
dé (2)
thé (2)
damner (2)
tas (2)
dot (2)
tort (2)

dos (2)
tôt (2)
doux (2)
tout (2)
deux (2)
teuton (2)
du (2)
tu (2)

Training Series, Sentences

Claude (2)
je (2)
veux (2)
Claude je veux (2)
un (2)
sty (2)
lo (2)
un stylo (2)

A (2)
bas (2)
con (2)
A bas, con (2)
sul (2)
a (2)
ffreux (2)
sul affreux (2)

Claude je (2)
Claude je veux (2)
un sty (2)
un stylo (2)
Claude je veux (2)
un stylo (2)
Claude je veux (2)
un stylo (2)
Claude je veux (2)
un stylo (2)

A bas, (2)
A bas, con (2)
sul a (2)
sul affreux (2)
A bas, con (2)
sul affreux (2)
A bas, con (2)
sul affreux (2)
A bas, con (2)
sul affreux (2)

Post-test Warm-up

Claude (2)
Claude je veux (2)
Claude je veux un (2)
Claude je veux un sty (2)
Claude je veux un stylo (2)

A bas (2)
A bas, con (2)
A bas, consul (2)
A bas, consul a (2)
A bas, consul affreux! (2)

The following shows the warm-up words and sentences for each day. Numbers preceding the sentences indicate their number in the Aptitude-Criterion Test. The target phonemes are underlined.

First day

Warm-up words: dis, tir, dé, thé, damner, tas, dot, tort, dos, tôt, doux, tout, deux, teuton, du, tu

Sentences: (9) Claude je veux un stylo.
(13) A bas, consul affreux!

Second day

Warm-up words: bis, pis, béret, paix, bas, pas, botte, porte, beau, paume, doux, pou, boeufs, peu, bu, du

Sentences: (5) Donne-moi sept tapis mauves!
(17) On a vu neuf bûches jaunes.

Third day

Warm-up words: lit, riz, les, raie, la, rat, lotte, roc, lot, rôle, loup, roue, pleut, creuse, lu, rue

Sentences: (1) Un bon ami vient tôt.
(21) Ton oncle plonge jusqu'au fond.

Fourth day

Warm-up words: banc, gland, grand, tant, tante, bon, bondé, son, sombre, pin, gain, dinde, feindre, timbre, brun, Verdun

Sentences: (3) Ils sont un peu fanés.
(23) C'est combien ce chapeau?

Fifth day

Warm-up words: patte, tante, robe, rude, déjà, teinte, crever, neuf, dogue, pire, bien, coq, seul, dame, âne, peigne

Sentences: (7) Conduis du bon côté.
(15) L'ozone a sauve Paul.

Sixth day

Warm-up words: côte, tort, trompe, peur, humble, creuse, tape, rond, rosse, champ, tour, rive, nuage, poêle, zèle, Jean

Sentences: (19) Au printemps il pleut trop.
(11) Deux à onze livres cinquante.

TRAINING PROGRAMS, CONTINUATION EXPERIMENT
(Experiment 6)

The following shows the program for the first day, the same pattern was followed for all six days. The method of testing and scoring and the schedule of presentation was the same as for the Replication Experiment (q.v.). Four, instead of two sentences, were presented each day and the warm-up words were omitted.

The number (2) after an utterance indicates it was presented twice in succession. Target phonemes are underlined.

Pre-test and Post-test

Claude (2)
je (2)
veux (2)
Claude je
Claude je veux (2)
Claude je veux un
Claude je veux un sty
Claude je veux un stylo. (2)

Chan (2)
tons (2)
cette (2)
Chantons
Chantons cette (2)
Chantons cette belle
Chantons cette belle chan
Chantons cette belle chanson. (2)

A (2)
bas (2)
con (2)
A bas
A bas, con (2)
A bas, consul
A bas, consul a
A bas, consul affreux! (2)

Il (2)
a (2)
su (2)
Il a
Il a su (2)
Il a su le
Il a su le faire
Il a su le faire seule. (2)

Training Series

Claude (2)
je (2)
veux (2)
Claude je veux (2)
un (2)
sty (2)
lo (2)
un stylo (2)
Claude je veux (2)
un stylo (2)

Chan (2)
tons (2)
cette (2)
Chantons cette (2)
belle (2)
chan (2)
son (2)
belle chanson (2)
Chantons cette (2)
belle chanson (2)

A (2)
bas (2)
con (2)
A bas, con (2)
sul (2)
a (2)
ffreux (2)
sul affreux (2)
A bas, con (2)
sul affreux (2)

Il (2)
a (2)
su (2)
Il a su (2)
le (2)
faire (2)
seule (2)
le faire seule (2)
Il a su (2)
le faire seule (2)

Claude je veux
un stylo
Claude je veux
un stylo

Chantons cette
belle chanson
Chantons cette
belle chanson

A bas, con
sul affreux
A bas, con
sul affreux

Il a su
le faire seule
Il a su
le faire seule

Claude je veux
un stylo

Chantons cette
belle chanson

A bas, con
sul affreux

Il a su
le faire seule

Warm-up for Post-test

Claude (2)
Claude je veux (2)
Claude je veux un (2)
Claude je veux un sty (2)
Claude je veux un stylo (2)

Chantons (2)
Chantons cette (2)
Chantons cette belle (2)
Chantons cette belle chan (2)
Chantons cette belle chanson (2)

A bas (2)
A bas, con (2)
A bas, consul (2)
A bas, consul a (2)
A bas, consul affreux!(2)

Il a (2)
Il a su (2)
Il a su le (2)
Il a su le faire (2)
Il a su le faire seule (2)

The following lists the sentences for each day. Numbers preceding the sentences indicate their number in the Aptitude-Criterion Test. The target phonemes are underlined.

First day: (9) Claude je veux un stylo.
(13) A bas, consul affreux!
(18) Chantons cette belle chanson.
(20) Il a su le faire seule.

Second day: (5) Donne moi sept tapis mauves.
(17) On a vu neuf bûches jaunes.
(2) Sur les monts chiens et daims.
(12) C'est la plainte d'aucune sainte.

Third day: (1) Un bon ami vient tôt.
(21) Ton oncle plonge jusqu'au fond.
(16) Julie veut faire la queue.
(14) Le gamin fin l'atteint.

Fourth day: (3) Ils sont un peu fanés.
(23) C'est combien ce chapeau.
(24) Jules monte et tombe cinq fois.
(22) Jouons bien au ton d'un luth.

Fifth day: (7) Conduis du bon côté.
(15) L'ozone a sauve Paul.
(4) Ôte ce bon preux de l'eau.
(10) Sa mule peureuse culbute.

Sixth day: (19) Au printemps il pleut trop.
(11) Deux à onze livres cinquante.
(6) Une jument heurte le sol.
(8) Tu as les cheveux plats.

APTITUDE-CRITERION TEST, TRIAL EXPERIMENT
(Experiment 3)

The following shows the two patterns of build-up used in this test, type A and type B. Utterances followed by (2) were repeated a second time. Slant marks indicate a brief pause between syllables. For Type A, the second presentations of the second and final utterances in the build-up were scored for approximation of the whole utterance to the French phonological pattern. For Type B, the fifth and final utterances were so scored. For both types the two underlined phonemes were scored in the second presentation of the final utterance.

Type A Build-up	Type B Build-up
Joue la (2)	Le (2)
Joue la reine (2)	Dos (2)
Joue la reine de (2)	Du (2)
Joue la reine de coeur. (2)	Le / dos / du (2)
	Le dos du
	Le dos du beau (2)
	Le dos du beau bebé. (2)

The following are the sentences used in the test. The letter preceding an utterance indicates the type of build-up for that utterance. The numbers preceding scored utterances indicate their order in the test. The target phonemes are underlined.

First Half (Presented with Activated Headphones)

Warm-up sentences, not scored:

(A) Ami, qui joue ici?
(A) Un sou pour ces joujoux.
(A) Voilà un gros pot d'eau.
(B) Celle que j'aime c'est sa soeur.
(B) Ton teint est très laiteux.
(B) Trois enfants font des bonds.

Scored sentences:

(1-B) Le dos du beau bébé.
(2-B) Les cimes des monts sont hautes.
(3-B) Le chou rouge et le riz.
(4-B) Ces deux jeux sont fameux.
(5-B) Ce nain se met dans le coin.
(6-B) Le chien blanc a bien faim.
(7-A) Joue la reine de coeur.
(8-A) La moto arrive tôt.
(9-A) Il court dans la rue.
(10-A) Une fleur jaune n'est pas belle.
(11-A) Il est fort comme un boeuf.
(12-A) Il tombe sur le menton.

Second Half (Presented with Inactivated Headphones)

Warm-up sentences, not scored:

(A) Paul joue à la pelotte.
(A) Laissez ces p'tits bébés.
(A) Il aime bien le bon vin.
(B) Lucien avait bien faim.
(B) Il n'ya rien dans deux coins.
(B) Voici le hibou rouge.

Scored sentences:

(13-B) Buvez du thé chaud.
(14-B) Donnez-nous des beaux mots.
(15-B) Charlot veut faire la queue.
(16-B) Hélène a peur du boeuf.
(17-B) Bois un verre de liqueur.
(18-B) La dame blonde danse toute seule.

(19-A) Un morceau de gâteau.
(20-A) Qui veut un bon café?
(21-A) L'enfant chante une chanson.
(22-A) Il a pu voir la lune.
(23-A) Les grosses pommes dans le seau.
(24-A) Cinq rats trouvent du pain noir.

TRAINING PROGRAMS, TRIAL EXPERIMENT
(Experiment 3)

The following shows the program for the first day. The same pattern was followed in all succeeding days. The entire program, including the Pre-tests was presented twice with an eight-minute intermission between presentations. Then the Review and Post-test was presented. The Pre-tests were scored for the first presentation. An (S) preceding an utterance indicates that it was scored for approximation of the whole utterance to the French phonological pattern. The scored phonemes are underlined.

(Read both columns to horizontal line)

Pre-test, First Utterance

les
cimes
des
monts
sont
hautes
les
cimes
des

monts
sont
hautes
les / cimes / des
monts / sont / hautes
(S) les cimes des
les cimes des monts (2)
(S) les cimes des monts sont hautes (2)

Build-up, First Utterance

les (2)
cimes (2)
des (2)
les / cimes / des (2)
monts (2)
sont (2)
hautes (2)
monts / sont / hautes (2)

les / cimes / des
monts / sont / hautes
les / cimes / des
monts / sont / hautes
les cimes des (2)
les cimes des monts (2)
les cimes des monts sont hautes(2)

Pre-test, Second Utterance

la
mo
to
a
rrive
tôt
la
mo
to

a
rrive
tôt
la / mo / to
a / rrive / tôt
la / mo / to
(S) la moto
la moto arrive (2)
(S) la moto arrive tôt (2)

Build-up, Second Utterance

la (2)
mo (2)
to (2)
la / mo / to (2)
a (2)
rrive (2)
tôt (2)
a/ rrive / tôt (2)

la / mo / to
a / rrive / tôt
la / mo / to
a / rrive / tôt
la moto (2)
la moto arrive (2)
la moto arrive tôt (2)

Build-up, Alternating Utterances

les / cimes / des
les cimes des (2)
les cimes des monts (2)
les cimes des monts sont (2)
les cimes des monts sont hautes (2)

la / mo / to
la moto (2)
la moto arrive (2)
la moto arrive tôt (2)

les cimes des
les cimes des monts
les cimes des monts sont
les cimes des monts sont hautes

la moto
la moto arrive
la moto arrive tôt

les cimes des monts
les cimes des monts sont hautes
la moto arrive
la moto arrive tôt

les cimes des monts sont hautes
la moto arrive tôt
les cimes des monts sont hautes
la moto arrive tôt
les cimes des monts sont hautes
la moto arrive tôt

Review and Post-test

les / cimes / des (2)	(S)	les cimes des
monts / sont / hautes (2)		les cimes des monts
		les cimes des monts sont hautes
les / cimes / des		
monts / sont / hautes	(S)	la moto
les / cimes / des		la moto arrive
monts / sont / hautes		la moto arrive tôt
la / mo / to (2)		les cimes des monts sont hautes
a / rrive / tôt (2)		la moto arrive tôt
		les cimes des monts sont hautes
la / mo / to		la moto arrive tôt
a / rrive / tôt	(S)	les cimes des monts sont hautes
la / mo / to	(S)	la moto arrive tôt
a / rrive / tôt		

The following shows the sentences for each day. Numbers preceding sentences indicate their number in the Aptitude-Criterion Test. The target phonemes are underlined.

First Day: (2) Les cimes des monts sont hautes.
(8) La moto arrive tôt.

Second Day: (14) Donnez-nous des beaux mots.
(20) Qui veut un bon café?

Third Day: (4) Ces deux jeux sont fameux.
(10) Une fleur jaune n'est pas belle.

Fourth Day: (16) Hélène a peur du boeuf.
(22) Il a pu voir la lune.

Fifth Day: (6) Le chien blanc a bien faim.
(12) Il tombe sur le menton.

Sixth Day: (18) La dame blonde danse toute seule.
(24) Cinq rats trouvent du pain noir.

APPENDIX II: QUESTIONNAIRE

In the following questions circle the letter before the answer which comes closest to your feeling or belief:

1. I was in the following group
   a. Short delay
   b. Activated
   c. Inactivated
   d. Long delay

2. What I learned in this experiment
   a. Will never be of any value to me.
   b. May be of some value to me.
   c. Will definitely be of value to me.

3. a. All the work I did in this experiment was interesting.
   b. Some of the work I did in this experiment was interesting.
   c. None of the work I did was interesting.

4. a. None of the work was boring.
   b. Some of the work was boring.
   c. Some of the work was very boring.

5. a. I did my best almost all of the time.
   b. I did my best more than half the time.
   c. I did my best less than half the time.
   d. I wasn't really trying any of the time.

Write brief answers to the following questions:

1. What parts of the experiment were most interesting?

2. What parts of the experiment bored you?

3. What things about the experiment irritated you?

4. What bothered you so that you couldn't do your best work? How important was this?

5. How would you advise us to change the way we went about the experiment?

6. What did you like about the way we went about the experiment?

APPENDIX III

ANALYSIS OF RESPONSES TO OPEN-ENDED ITEMS ON QUESTIONNAIRE BY EXPERIMENT AND TREATMENT

		Ex4	Ex5	Ex6	Ex7	Total
Item 1: What interesting?						
A. Testing (usually pre-post-tests) or observing own progress	IF	3	4	4	2	13
	AF	1	1	2	2	6
	LD	3	1	1	1	6
	SD	2	5	2	-	9
	Total	9	11	9	5	34
B. Playback (Ss in IF and AF were allowed to hear their recording played back after the final test)	IF	2	-	-	1	3
	AF	2	-	-	-	2
	LD	5	1	2	3	11
	SD	3	1	2	4	10
	Total	12	2	4	8	26
C. Classroom work	IF	3	1	-	-	4
	AF	4	-	1	-	5
	LD	1	-	-	1	2
	SD	1	-	-	1	2
	Total	9	1	1	2	13
D. Learning new sounds or pronunciation of new language	IF	-	1	3	1	5
	AF	-	1	1	1	3
	LD	-	1	-	-	1
	SD	1	-	-	1	2
	Total	1	3	4	3	11
E. Everything was interesting	IF	-	1	-	1	2
	AF	-	1	-	2	3
	LD	-	-	-	-	-
	SD	1	-	-	1	2
	Total	1	2	-	4	7
F. Nothing was interesting or no part more interesting than another	IF	-	1	-	-	1
	AF	-	-	1	-	1
	LD	1	2	2	-	5
	SD	-	-	-	-	-
	Total	1	3	3	-	7
G. Practice or training sessions	IF	2	-	-	-	2
	AF	-	1	-	-	1
	LD	-	-	-	1	1
	SD	1	-	-	1	2
	Total	3	1	-	2	6

		Ex4	Ex5	Ex6	Ex7	Total
H. Miscellaneous	IF	2	-	-	3	5
	AF	-	5	4	1	10
	LD	-	3	3	-	6
	SD	1	2	2	-	5
	Total	3	10	9	4	26
Item 2: What was boring?						
A. Repetitiousness	IF	3	3	4	2	12
	AF	1	2	2	2	7
	LD	2	6	5	2	15
	SD	1	5	1	1	8
	Total	7	16	12	7	42
B. Second practice session (Playback for LD) or length of practice	IF	2	1	1	2	6
	AF	1	1	2	1	5
	LD	1	1	1	3	6
	SD	2	-	-	1	3
	Total	6	3	4	7	20
C. Nothing was boring	IF	1	-	1	-	2
	AF	-	-	1	3	4
	LD	3	1	-	2	6
	SD	2	-	-	4	6
	Total	6	1	2	9	18
D. Practice Sessions (without qualification)	IF	2	2	2	2	8
	AF	2	-	-	-	2
	LD	1	-	1	1	3
	SD	-	-	4	-	4
	Total	5	2	7	3	17
E. Various parts of instructions, testing machines, warm-ups for tests	IF	-	1	-	-	1
	AF	1	-	1	-	2
	LD	-	-	-	-	-
	SD	1	-	2	-	3
	Total	2	1	3	-	6
F. Miscellaneous	IF	2	-	-	2	4
	AF	1	4	-	-	5
	LD	3	-	-	1	4
	SD	2	2	-	-	4
	Total	8	6	-	3	17
Item 3: What irritated?						
A. Nothing irritated	IF	3	3	2	2	9
	AF	3	3	2	6	14
	LD	-	1	3	4	8
	SD	-	1	2	5	8
	Total	6	7	9	17	39

		Ex4	Ex5	Ex6	Ex7	Total
B. Various features of instructions and testing equipment	IF	3	3	-	-	6
	AF	1	3	3	-	7
	LD	1	2	1	-	4
	SD	-	7	-	-	7
	Total	5	15	4	-	24
C. Mechanical failures or variations in loudness	IF	-	-	-	-	-
	AF	1	-	-	-	1
	LD	1	-	-	1	2
	SD	8	-	-	1	9
	Total	10	-	-	2	12
D. Various causes of boredom	IF	-	1	2	-	3
	AF	1	-	1	-	2
	LD	-	2	1	-	3
	SD	-	-	3	-	3
	Total	1	3	7	-	11
E. Not knowing the meaning of the sentences	IF	2	-	-	-	2
	AF	-	-	-	-	-
	LD	5	-	-	-	5
	SD	1	-	-	-	1
	Total	8	-	-	-	8
F. Subject's own failures	IF	1	-	-	3	4
	AF	1	-	-	-	1
	LD	-	-	-	1	1
	SD	-	-	-	1	1
	Total	2	-	-	5	7
G. Delays: Waiting for latecomers, rest periods	IF	-	1	2	-	3
	AF	-	-	-	-	-
	LD	-	1	-	-	1
	SD	-	2	-	-	2
	Total	-	4	2	-	6
H. Uncomfortable headphones	IF	-	-	-	-	-
	AF	-	1	1	1	3
	LD	-	2	-	-	2
	SD	-	-	-	-	-
	Total	-	3	1	1	5
I. Miscellaneous	IF	-	-	1	2	3
	AF	1	-	1	-	2
	LD	1	2	1	1	5
	SD	-	1	-	-	1
	Total	2	3	3	3	11

		Ex4	Ex5	Ex6	Ex7	Total
Item 4: What bothered?						
A. Nothing bothered so as to prevent best work	IF	3	1	-	1	5
	AF	1	3	4	3	11
	LD	3	3	6	4	16
	SD	1	1	2	2	6
	Total	8	8	12	10	38
B. External disturbances, others voices, movements of proctors, self-consciousness when others heard	IF	1	1	2	2	6
	AF	4	1	-	-	5
	LD	3	1	-	2	6
	SD	4	1	-	-	5
	Total	12	4	2	4	22
C. Difficult material, remembering long sentences, discriminating sounds, frustration at failure	IF	1	1	-	3	5
	AF	-	1	1	3	5
	LD	-	1	-	2	3
	SD	-	1	-	2	3
	Total	1	4	1	10	16
D. Repetitiousness, boredom, mind wandering, losing interest	IF	-	1	3	-	4
	AF	2	2	-	-	4
	LD	-	1	1	-	2
	SD	-	1	2	1	4
	Total	2	5	6	1	14
E. Fatigue, drowsiness, yawning	IF	-	2	2	-	4
	AF	-	-	1	-	1
	LD	1	-	-	-	1
	SD	-	2	2	-	4
	Total	1	4	5	-	10
F. Equipment breakdown or malfunction	IF	1	-	-	-	1
	AF	1	-	-	-	1
	LD	1	-	-	1	2
	SD	2	-	-	-	2
	Total	5	-	-	1	6
G. Miscellaneous	IF	2	1	-	1	4
	AF	-	-	-	1	1
	LD	-	-	-	1	1
	SD	1	-	-	-	1
	Total	3	1	-	3	7
H. Difficulty specified as being "important", "fairly important", "important when it happened"	IF	-	1	1	1	3
	AF	-	-	-	1	1
	LD	-	-	-	-	-
	SD	1	-	1	1	3
	Total	1	1	2	3	7

		Ex4	Ex5	Ex6	Ex7	Total
I. Difficulty specified as being not important	IF	1	-	1	4	6
	AF	1	2	1	-	4
	LD	3	-	-	-	3
	SD	2	-	2	3	7
	Total	7	2	4	7	20
Item 5: Suggested changes						
A. No suggestion or good as it is	IF	2	2	3	4	11
	AF	3	3	4	4	14
	LD	2	4	4	3	13
	SD	4	-	-	4	8
	Total	11	9	11	15	46
B. Give meanings or show words visually	IF	4	1	1	-	6
	AF	-	-	-	-	-
	LD	3	-	-	-	3
	SD	2	-	-	-	2
	Total	9	1	1	-	11
C. Suggestions for improvement of equipment and facilities	IF	1	2	2	1	6
	AF	2	-	-	1	3
	LD	-	-	-	-	-
	SD	-	-	-	-	-
	Total	3	2	2	2	9
D. Decrease repetitiousness of practice material	IF	-	1	2	-	3
	AF	-	-	-	-	-
	LD	-	1	1	1	3
	SD	-	2	-	-	2
	Total	-	4	3	1	8
E. Omit or change classroom work or (in Ex6) include classroom or perform functions of classroom	IF	-	2	-	-	2
	AF	-	1	-	-	1
	LD	-	-	1	-	1
	SD	-	2	2	-	4
	Total	-	5	3	-	8
F. Make instructions less repetitious, less "child-like" or let students start machines	IF	-	-	-	-	-
	AF	-	2	1	-	3
	LD	1	1	-	-	2
	SD	-	2	-	-	2
	Total	1	5	1	-	7
G. Miscellaneous	IF	1	1	1	2	5
	AF	1	1	1	1	4
	LD	2	2	2	3	9
	SD	2	4	2	3	11
	Total	6	8	6	9	29

		Ex4	Ex5	Ex6	Ex7	Total
Item 6: What like?						
A. Efficiency, good organization, good planning, no time wasted	IF	1	4	5	1	11
	AF	1	2	3	1	7
	LD	-	3	3	-	6
	SD	3	3	3	-	9
	Total	5	12	14	2	33
B. 8 minute rest pause	IF	3	-	-	1	4
	AF	2	1	-	1	4
	LD	2	-	1	2	5
	SD	4	-	-	1	5
	Total	11	1	1	5	18
C. Friendliness, cheerfulness, helpfulness of staff	IF	1	1	1	3	6
	AF	2	1	1	-	4
	LD	2	-	-	1	3
	SD	2	-	1	1	4
	Total	7	2	3	5	17
D. Classroom work	IF	2	-	-	3	5
	AF	2	-	-	-	2
	LD	3	-	-	-	3
	SD	2	1	-	1	4
	Total	9	1	-	4	14
E. Laboratory sessions or practice sessions or tests or working with machines	IF	3	-	-	-	3
	AF	-	-	-	2	2
	LD	1	-	-	2	3
	SD	2	1	-	1	4
	Total	6	1	-	5	12
F. Blank or "nothing in particular"	IF	-	1	1	-	2
	AF	-	1	-	1	2
	LD	-	4	2	-	6
	SD	-	2	-	1	3
	Total	-	8	3	2	13
G. Everything	IF	-	-	-	-	-
	AF	1	-	1	1	3
	LD	-	-	-	-	-
	SD	2	1	1	-	4
	Total	3	1	2	1	7
H. Playback or hearing own voice	IF	1	-	-	-	1
	AF	-	-	-	-	-
	LD	2	-	-	2	4
	SD	-	-	2	-	2
	Total	3	-	2	2	7

		Ex4	Ex5	Ex6	Ex7	Total
I. Was interesting or "made interesting"	IF	-	-	-	-	-
	AF	1	-	1	1	3
	LD	-	-	-	-	-
	SD	2	1	1	-	4
	Total	1	1	1	3	6
K. Miscellaneous	IF	1	1	-	3	5
	AF	2	2	3	2	9
	LD	1	-	1	2	4
	SD	4	-	4	1	9
	Total	8	3	8	8	27

APPENDIX IV

SUPPLEMENTARY TABLES

Contents

Page

TABLE A -------------------------- 101

TABLE B -------------------------- 102

TABLE C -------------------------- 102

TABLE D -------------------------- 103

TABLE E -------------------------- 104

TABLE F -------------------------- 105

TABLE G -------------------------- 105

TABLE H -------------------------- 106

TABLE I -------------------------- 107

TABLE J -------------------------- 108

TABLE K -------------------------- 109

TABLE L -------------------------- 109

TABLE M -------------------------- 110

TABLE A

RELIABILITY COEFFICIENTS DERIVED FROM ANALYSIS OF THE APTITUDE TEST FOR ALL THREE REPLICATION EXPERIMENTS

(Underlined split-half coefficients are corrections by the Spearman-Brown prophecy formula. Underlined coefficients for correlations between raters are corrections for attenuation by the Spearman formula.)

	Ex4	Ex5	Ex7
N =	28	27	28
Split-half coefficients, individual raters:			
	.85	.54	.77
Rater C, Phonemic	.92	.70	.86
	.78	.73	.65
Rater S, Phonemic	.88	.84	.79
	.90	.88	.91
Rater C, Overall	.95	.94	.95
	.87	.78	.80
Rater B, Overall	.93	.88	.89
Split-half coefficients, both rater's scores combined:			
	.86	.70	.77
Phonemic	.94	.83	.86
	.92	.85	.92
Overall	.96	.92	.96
	.93	.84	.89
Combined Phonemic & Overall scores	.96	.91	.94
Correlations between raters:			
Rater C x Rater S, Phonemic	.89	.81	.82
	.99	1.05	1.00
Rater C x Rater B, Overall	.93	.87	.88
	.99	.96	.96

TABLE B

MEAN SCORES OF RATERS ON APTITUDE, FINAL TRAINED AND FINAL UNTRAINED, WITH DIFFERENCES BETWEEN RATERS AND SELF-DIFFERENCES BETWEEN MEANS OF ODD AND EVEN ITEMS ON THE APTITUDE TEST.

Special abbreviations: C, S, and B; Raters C, S, and B.
D; differences between raters or self-differences.

MEANS AND DIFFERENCES BETWEEN RATERS

		Experiment 4			Experiment 5			Experiment 7		
		Ap	FT	FU	Ap	FT	FU	Ap	FT	FU
	C	690	827	823	704	814	807	588	768	732
Ph	S	634	744	704	611	705	665	567	684	678
	D	56	83	119	93	109	142	21	84	54
	B	792	994	883	941	1062	1028	921	1085	1018
OA	C	615	856	754	679	831	780	568	807	713
	D	177	138	129	262	231	248	353	278	305

MEANS ON ODD AND EVEN ITEMS WITH SELF-DIFFERENCES

		Ex4		Ex5		Ex7	
		C	S	C	S	C	S
	ODD	691	640	706	628	604	559
Ph	EVEN	690	627	702	593	572	575
	D	1	13	4	35	32	-16
		C	B	C	B	C	B
	ODD	620	802	683	933	584	921
OA	EVEN	610	781	675	950	552	921
	D	10	21	8	-17	32	0

TABLE C

CORRELATIONS BETWEEN PHONEMIC AND OVERALL VARIABLES

(Underlined coefficients are corrections by Spearman-Brown prophesy formula to compare criterion test intercorrelations with Aptitude Test correlations.)

	N	Ap	Pr		Po		FT		FU	
Ex4	28	.85	.83	.91	.90	.95	.91	.95	.86	.94
Ex5	27	.73	.71	.83	.57	.73	.48	.65	.64	.78
Ex7	28	.64	.63	.77	.77	.86	.72	.84	.55	.71

TABLE D

INTERCORRELATIONS BETWEEN TESTS IN THE REPLICATION EXPERIMENT FOR BOTH OVERALL AND PHONEMIC VARIABLES WITH MEANS OF CORRELATIONS.

		Experiment 4				Experiment 5				Experiment 7			
		Pr	Po	FT	FU	Pr	Po	FT	FU	Pr	Po	FT	FU
	Ap	.77	.80	.78	.85	.38	.34	.49	.61	.72	.70	.74	.75
Ph-	Pr		.85	.86	.87		.77	.53	.48		.67	.77	.71
Ph	Po			.88	.89			.56	.42			.87	.60
	FT				.85				.80				.74
	Ap	.84	.80	.76	.89	.64	.43	.53	.81	.74	.53	.66	.85
OA-	Pr		.90	.83	.88		.83	.75	.66		.79	.83	.84
OA	Po			.89	.85			.87	.63			.78	.68
	FT				.85				.67				.81
	Ap	.80	.85	.84	.87	.50	.37	.43	.64	.68	.62	.73	.71
Ph-	Pr		.86	.84	.77		.57	.32	.22		.73	.69	.74
OA	Po			.90	.81			.36	.17			.61	.53
	FT				.79				.51				.66
	Ap	.66	.67	.74	.79	.16	.24	.44	.46	.58	.41	.57	.43
OA-	Pr		.80	.77	.86		.58	.62	.63		.55	.67	.60
Ph	Po			.86	.90			.67	.68			.72	.62
	FT				.87				.63				.68

Mean Correlation Coefficients, All Three Experiments

(Computed through z-transformations)

Special Abbreviation: Tr; training criterion tests, Pr, Po, FT.

	PhTr		OATr		CoTr		PhFU		OAFU		CoFU	
	N	M	N	M	N	M	N	M	N	M	N	M
PhAp	9	.66	9	.67	18	.67	3	.75	3	.76	6	.76
OAAp	9	.52	9	.68	18	.61	3	.59	3	.77	6	.69
CoAp	18	.60	18	.68	36	.64	6	.68	6	.77	12	.73
PhTr	9	.78	9	.70	18	.74	9	.74	9	.62	18	.68
OATr	9	.71	9	.84	18	.78	9	.75	9	.78	18	.76
CoTr	18	.74	18	.78	36	.76	18	.74	18	.71	36	.73

TABLE E

MEANS OF TREATMENTS IN THE REPLICATION EXPERIMENTS FOR THE PHONEMIC AND OVER-ALL VARIABLES WITH DIFFERENCES BETWEEN APTITUDE AND CRITERION SCORES AND WITH MEANS OF EACH EXPERIMENT.

Special abbreviations: DPr, DPo, DFT, DFU; differences between Aptitude Test and criterion test scores.

			N	Ap	Pr	Po	FT	FU	DPr	DPo	DFT	DFU
P		IF	7	625	768	800	779	757	143	175	154	132
	Ex	AF	7	701	843	880	854	809	142	179	153	108
H	4	LD	7	674	776	801	781	765	82	127	107	91
		SD	7	649	704	763	729	723	55	114	80	74
O		Mm		662	773	811	786	763	111	149	124	101
		IF	7	662	717	767	723	737	55	105	61	75
N	Ex	AF	7	653	793	842	774	747	140	189	121	94
	5	LD	7	665	776	824	762	743	111	159	97	78
E		SD	6	648	746	832	781	713	98	184	133	65
		Mm		657	758	816	760	735	101	159	103	78
M		IF	7	557	664	733	705	688	107	176	148	131
	Ex	AF	7	538	692	731	700	677	154	193	162	139
I	7	LD	7	590	715	772	768	721	125	182	178	131
		SD	7	625	703	712	730	733	78	87	105	108
C		Mm		578	694	737	726	705	116	159	148	127
O		IF	7	691	815	887	902	815	124	196	211	124
	Ex	AF	7	701	863	954	974	838	162	253	273	137
V	4	LD	7	732	826	893	939	824	94	161	207	92
		SD	7	688	809	852	885	795	121	164	197	107
E		Mm		703	828	897	925	818	125	194	222	115
		IF	7	829	878	938	926	931	49	109	97	102
R	Ex	AF	7	777	911	988	954	892	134	211	177	115
	5	LD	7	798	898	976	949	898	100	178	151	100
A		SD	6	838	919	954	958	895	81	116	120	57
		Mm		810	901	964	947	904	91	154	137	94
L		IF	7	752	872	936	951	871	120	184	199	119
	Ex	AF	7	699	835	915	905	820	136	216	206	121
L	7	LD	7	745	884	960	958	890	139	215	213	145
		SD	7	781	880	929	969	881	99	148	188	100
		Mm		744	868	935	946	865	124	191	202	121

TABLE F

MEAN GAINS ON THE PRE-TESTS BY DAYS ABOVE SCORES ON THE SAME ITEMS ON THE APTITUDE TEST.

		N	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
Ph	Ex4	28	106	206	80	43	125	80
	Ex5	27	182	82	97	41	99	45
	Ex7	28	120	136	129	136	54	98
	Mm		136	142	102	73	92	75
OA	Ex4	28	33	131	120	149	94	176
	Ex5	27	121	80	115	40	68	131
	Ex7	28	105	117	138	188	52	92
	Mm		86	110	124	127	71	133
Co	Mm		111	126	113	100	82	104

TABLE G

MEAN PHONEMIC, OVERALL, AND COMBINED SCORES BY TREATMENTS FOR THE FIRST THREE AND LAST THREE DAYS ON THE PRE-TESTS WITH DIFFERENCES BETWEEN THEM.

Special Abbreviations: 1st; mean of first three Pre-test days. 2nd; mean of last three Pre-test days. Dif; difference between 1st and 2nd.

	Phonemic				Overall				Combined			
	IF	AF	LD	SD	IF	AF	LD	SD	IF	AF	LD	SD
1st	709	789	750	717	854	865	967	860	782	927	809	789
2nd	723	763	761	715	856	877	871	873	789	820	816	794
Dif	-14	+26	-11	+2	-2	-12	-4	-13	-8	+7	-7	-5

TABLE H

ADJUSTED MEANS OF TREATMENTS DERIVED FROM ONE-WAY ANALYSIS OF COVARIANCE OF EACH CRITERION TEST IN EACH EXPERIMENT FOR PHONEMIC AND OVERALL VARIABLES IN THE REPLICATION EXPERIMENTS.

		Phonemic			Overall		
		Ex4	Ex5	Ex7	Ex4	Ex5	Ex7
Pre-test	IF	798	715	683	825	866	865
	AF	802	795	729	864	930	870
	LD	760	773	703	802	904	885
	SD	711	750	658	821	903	852
Post-test	IF	841	765	755	898	928	933
	AF	838	844	773	956	1005	936
	LD	788	821	759	868	981	960
	SD	778	836	663	866	940	913
Final Trained	IF	815	720	725	912	915	947
	AF	818	777	738	975	970	923
	LD	770	758	756	917	956	958
	SD	741	786	685	897	945	954
Final Untrained	IF	793	735	702	826	918	865
	AF	771	750	704	839	914	855
	LD	754	738	713	800	905	890
	SD	737	720	702	808	876	853

TABLE I

TWO-WAY ANALYSES OF COVARIANCE OF IF, AF AND LD TREATMENTS BY EXPERIMENTS 4, 5, and 7 FOR PHONEMIC AND OVERALL SCORES ON ALL FOUR CRITERION TESTS.

PHONEMIC

Test	Source	df	MS	F	Sig.
Pre-tests	Treatments	2	729	3.59	$P < .05$
	Experiments	2	319	1.57	N.S.
	Interaction	4	205	1.01	N.S.
	Within cells	53	203		
Post-tests	Treatments	2	465	1.93	N.S.
	Experiments	2	129	.54	N.S.
	Interaction	4	303	1.26	N.S.
	Within cells	53	241		
Final Trained	Treatments	2	215	.91	N.S.
	Experiments	2	874	3.70	$P < .05$
	Interaction	4	267	1.13	N.S.
	Within cells	53	236		
Final Untrained	Treatments	2	46	.21	N.S.
	Experiments	2	387	1.80	N.S.
	Interaction	4	50	.23	N.S.
	Within cells	53	215		

OVERALL

Test	Source	df	MS	F	Sig.
Pre-tests	Treatments	2	404	2.48	N.S.
	Experiments	2	90	.55	N.S.
	Interaction	4	240	1.47	N.S.
	Within cells	53	163		
Post-tests	Treatments	2	761	4.48	$P < .05$
	Experiments	2	52	.31	N.S.
	Interaction	4	347	2.04	N.S.
	Within cells	53	170		
Final Trained	Treatments	2	411	2.46	N.S.
	Experiments	2	644	3.86	$P < .05$
	Interaction	4	256	1.53	N.S.
	Within cells	53	167		
Final Untrained	Treatments	2	31	.25	N.S.
	Experiments	2	82	.66	N.S.
	Interaction	4	140	1.12	N.S.
	Within cells	53	125		

TABLE J

ADJUSTED MEANS OF TREATMENTS AND EXPERIMENTS DERIVED FROM TWO-WAY ANALYSIS OF COVARIANCE FOR PHONEMIC AND OVERALL VARIABLES WITH MEANS OF MEANS AND DEVIATIONS FROM MEANS OF MEANS.

			Pr		Po		FT		FU	
			M	Dm	M	Dm	M	D	M	Dm
P		IF	725	-22	780	-15	748	-13	740	+2
H	Treat-	AF	773	+26	817	+22	774	+13	743	+5
O	ments	LD	743	-4	787	-8	760	-1	731	-7
N		Mm	747		795		761		738	
E		Ex4	768	+21	793	-1	775	+14	745	+7
M	Experi-	Ex5	737	-10	783	-11	728	-33	716	-22
I	ments	Ex7	737	-10	807	+13	779	+18	753	+15
C		Mm	747		794		761		738	
		IF	848	-17	915	-24	920	-20	863	-2
O	Treat-	AF	885	+20	965	+26	957	+17	870	+5
V	ments	LD	861	-4	936	-3	943	+3	861	-4
E		Mm	865		939		940		865	
R		Ex4	862	-4	934	-4	962	+22	861	-3
A	Experi-	Ex5	857	-7	935	-3	911	-29	858	-6
L	ments	Ex7	874	+10	946	+8	947	+7	874	+10
L		Mm	864		938		940		864	

TABLE K

TWO-WAY MULTIPLE REGRESSION ANALYSES OF COVARIANCE OF IF, AF, AND LD TREATMENTS BY EXPERIMENTS 4, 5, AND 7 FOR ALL FOUR CRITERION TESTS USING COMBINED PHONEMIC AND OVERALL SCORES.

Test	Source	df	MS	F	Sig.
	Treatments	2	1697.3	2.96	N.S.
Pre-tests	Experiments	2	1236.0	2.15	N.S.
	Interaction	4	257.3	.45	N.S.
	Within cells	52	574.6		
	Treatments	2	1939.8	3.18	P<.05
Post-tests	Experiments	2	281.2	.46	N.S.
	Interaction	4	1291.6	2.12	N.S.
	Within cells	52	610.1		
	Treatments	2	1117.6	2.10	N.S.
Final Trained	Experiments	2	3468.3	6.52	P<.01
	Interaction	4	805.6	1.52	N.S.
	Within cells	52	531.6		
	Treatments	2	253.7	.59	N.S.
Final Untrained	Experiments	2	1003.8	2.33	N.S.
	Interaction	4	378.1	.88	N.S.
	Within cells	52	431.6		

TABLE L

TWO-WAY ANALYSES OF COVARIANCE OF EXPERIMENTS 4, 5, AND 7 BY IF AND AF TREATMENTS AND ALSO BY AF AND LD TREATMENTS FOR THE POST-TESTS USING COMBINED PHONEMIC AND OVERALL SCORES.

Treatments	Source	df	MS	F	Sig.
	Treatments	2	3454.4	4.64	P<.05
AF and IF	Experiments	2	372.0	.50	N.S.
	Interaction	4	895.4	1.20	N.S.
	Within cells	34	744.3		
	Treatments	2	1648.1	2.23	N.S.
AF and LD	Experiments	2	392.4	.53	N.S.
	Interaction	4	587.1	.79	N.S.
	Within cells	34	739.5		

TABLE M

MEANS OF PHONEMIC AND OVERALL SCORES BY TREATMENTS FOR ALL TESTS IN EXPERIMENT 5, THE PART OF EXPERIMENT 6 TRAINED IN EXPERIMENT 5, AND THE PART OF EXPERIMENT 6 NOT TRAINED IN EXPERIMENT 5 WITH MEANS OF TREATMENT MEANS.

Note: In Experiment 6 "In" signifies Introductory Test and "F" Final Test.

		Experiment 5					Experiment 6 Trained in 5				Experiment 6 Untrained in 5			
		Ap	Pr	Po	FT	FU	In	Pr	Po	F	In	Pr	Po	F
Ph	IF	662	717	767	723	737	701	710	760	728	720	734	712	715
	AF	653	793	842	774	747	733	770	805	754	744	746	773	793
	LD	665	776	824	762	743	770	769	782	780	763	754	735	755
	SD	648	746	832	781	713	735	769	782	756	704	720	732	754
	Mm	657	758	816	760	735	735	754	782	754	734	739	738	755
OA	IF	829	878	938	926	931	899	929	947	947	896	905	949	943
	AF	777	911	988	954	892	945	946	960	969	863	902	961	928
	LD	798	898	976	949	898	940	948	945	950	887	901	936	938
	SD	838	919	954	958	895	943	955	939	974	879	884	949	953
	Mm	810	901	964	947	904	932	944	948	960	881	898	949	940